Chapter 3: Visualization of Structured Data

The greatest value of a picture is when it forces us to notice what we never expected to see–
John Tukey

Learning Objectives

- Understand how to use visualizations all through the data science process.
- Learn creating different types of charts to explore the data.
- Learn about Univariate and Multivariate exploratory analysis by using an
- example dataset.
- Understand visualization that would facilitate the model building and model
- evaluation processes.
- Learn to communicate the model to business and technical teams using visualization.
- Build and understand model interpretation using an example dataset.
- Build an interactive business operations dashboard.

Introduction

- Structured data (data in the matrix form) helps us understand what is happening in the organization.
- For example, revenue performance statistics, operational metrics, and so on.
- The characteristics of these data types drive which chart we can use for a given data on hand.

Introduction

Qualitative data can be further classified as:

- **Categorical** – Records that are grouped into multiple categories.
  - Example: Gender (male, female, and transgender), education level, student names, and state names.
- **Binary –** Variables with just two categories.
  - Example: Student result which can have values pass or fail, health diagnostic out- come which can be positive or negative.
- **Ordinal** – Variables with more than two categories and that have a natural order.
  - Example: Review ratings and Income level etc.

Introduction

Quantitative data can be further classified as:

- **Numerical Continuous** – variables that can take on any numeric value including fractional and decimal values.
  - Example: Market Share, sales amount, and height.
- **Numerical Discrete**–countably finite or infinite values for the records.
  - Example: Customers waiting at a bank, number of vehicles sold by a company, and number of students registered for a course.
- The characteristics of these data types drive which chart we can use for a given data on hand.

Introduction

Table 3.1 Datatypes and chart mapping

Variable 1	Variable 2	Chart Type	Comments
Numerical continuous data Ex: Height, weight, sales amount	-	Histogram We can also use - dot plots, boxplot	When we have single continuous data, we would like to understand the distribution of the values. Histograms help us to identify where does the most common values fall
Numerical continuous data Ex: Height, Sales Amount	Numerical continuous data Ex: Weight, profit amount	Scatter plot	Scatter plot shows the relationship (correlation) between two continuous variables
Numerical discrete data Ex: Count of bikes	-	Bar chart	Each bar represents a distinct value of bike count, and the height represents its proportion of this data in the entire sample
Categorical data Ex: State names	Numerical continuous data Ex: Profit Amount	Bar chart We can also use Pie chart if we have a smaller number of categories	Each bar represents a distinct value of state names, and the height represents its aggregated profit amount earned from the state such as sum and average
Binary data Ex: Student - pass or fail	-	Pie chart or bar chart	Each slice of pie or bar represents proportion of a binary data category in the entire sample
Ordinal data Ex: Review ratings	-	Bar chart	Each bar represents proportion of an ordinal data category in the entire sample

- Table 3.1 provides a simple mapping which we can use while deciding which basic charts can be used for which datatype.

Purchasing Data of Online Shoppers

- Let us use purchasing data of online shoppers available at the University of California Irvine (UCI) repository[1] to demonstrate various visuals.
- The data [online_shoppers_intention.csv] provided has information on the browse history of users who visited the website.
- The idea is to use this data to build a machine learning classification algorithm to predict a customer's intent to "buy" or "not buy".

[1]The data set has been provided by Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018) Web Address: http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset.
22020 Digital Trends report – https://www.adobe.com/in/offer/ digital-trends-2020.html

Data Description

Table 3.2 Purchase intention dataset description

Variable	Variable Type	Description
Administrative	Numeric	Number of administrative pages (login page, logout, register, account, forgot password etc.) visited by the user
Administrative_ Duration	Numeric	Total time (in seconds) This is the time spent by the visitor on administrative pages
Information	Numeric	Number of pages visited by the visitor on Information pages such as website, communication, and address information of the shopping site
Informational_ Duration	Numeric	Total amount of time (in seconds) spent by the visitor on informational pages
ProductRelated	Numeric	Number of pages visited by visitors about product related pages such as cart, product listing, product, and detail.
ProductRelated_ Duration	Numeric	Total amount of time (in seconds) spent by the visitor on product related page
BounceRates	Numeric	The percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session

(Continued)

Data Description

- The data consists of 12,330 sessions.
- Each session data belongs to a different user for a one-year period, which also avoids any **bias** towards specific days or campaigns.
- It contains 18 customer browsing features, out of which ten are numerical attributes and eight categorical attributes.
- The data description is provided in Table 3.2. The values of the features are:

  1. Details captured about pages browsed by the user. When the user is moving from one page to another in the e-commerce website, this information is derived from the Uniform Resource Locator (URL) in real time.

  2. All pages of the e-commerce site use Google Analytics to measure metrics.

Variable	Variable Type	Description
ExitRates	Numeric	The percentage of people who left a site from that page.
PageValues	Numeric	This is the average web pages visited by the user before completing an e-commerce transaction
SpecialDay	Numeric	This variable indicates if the user is visiting the site during peak selling days such as Mother's Day and Valentine's Day. This gives an indication that there is a higher probability of such sessions being converted into a transaction. For example, February 14 is Valentine's Day, so the variable will have non-zero value between February 2 and February 12. On February 8, it will have a maximum value of 1.
Month	Categorical	Month of the Visit date Data available for 10 months. No data for January and April
Operating_ Systems	Categorical	Operating system of the visitor. Numeric values 1 to 8 representing different operating systems
Browser	Categorical	Browser of the visitor Numeric values 1 to 13 representing different browser types
Region	Categorical	Geographic region from which the session has been started by the visitor. Numeric values 1 to 9 representing different regions
TrafficType	Categorical	Traffic source by which the visitor has arrived at the site (e.g., banner, SMS, direct). Numeric values 1 to 20 representing different traffic types
VisitorType	Categorical	Type of Visitor like "new visitor" and "returning visitor". 3 Categories: 1. Returning_Visitor 2. New_Visitor 3. Other
Weekend	Categorical	Boolean value indicating whether the date of the visit is weekend Boolean values (True or False)
Revenue	Categorical	Class label indicating whether the visit generate revenue or not Boolean values (True or False)

Exploratory Analysis

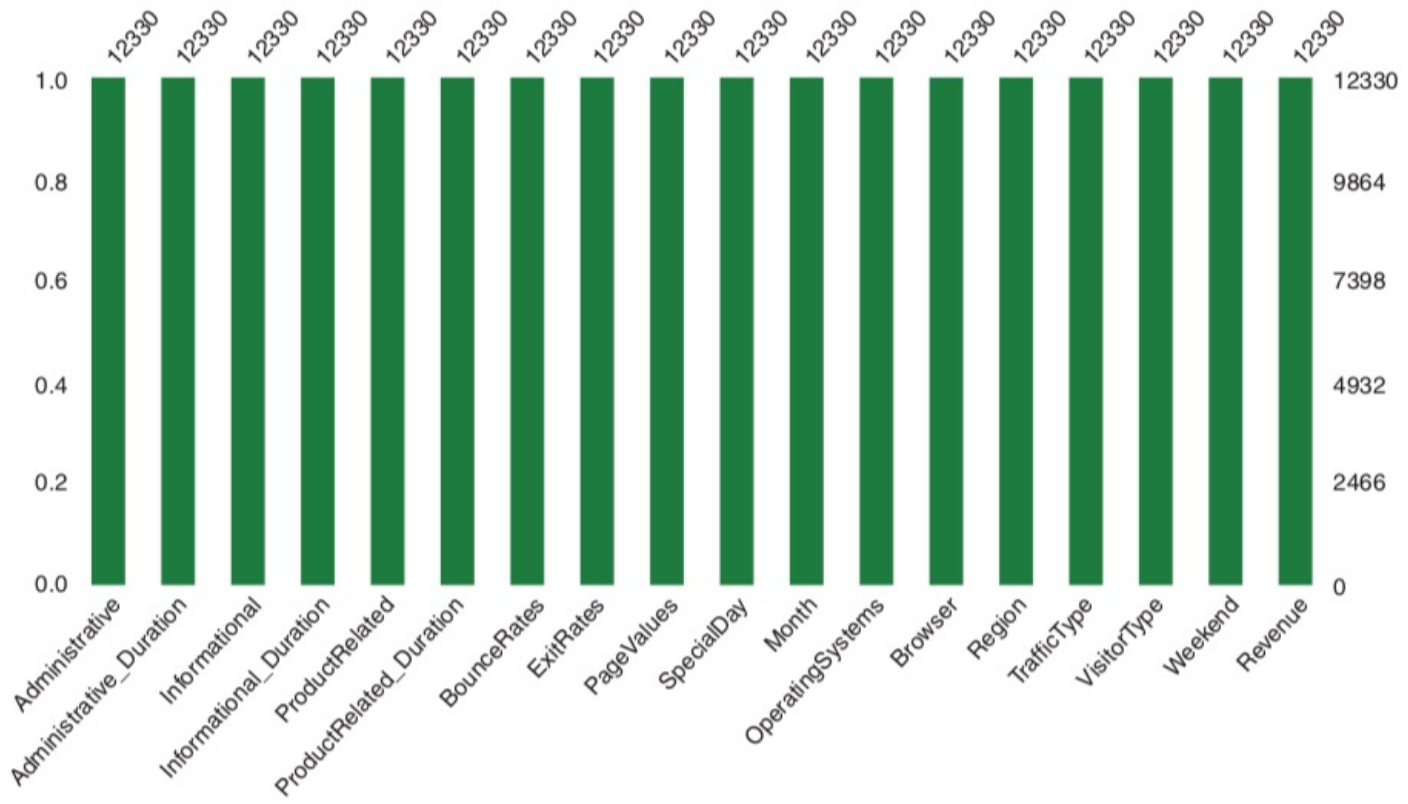


Figure 3.1 Bar chart for missing data check.

- From Fig. 3.1, we can conclude that there is no missing data across all the features in the provided dataset.

Univariate Analysis – How Do We Visualize Single Measure?

- Univariate analysis explores features in the dataset one variable at a time.
- In this section, we will learn and create the following chart types to infer insights for the provided data.
    1. Visualizing basic comparison charts
        (a) Bar chart
        (b) Big number
        (c) Histogram
        (d) Box and Whisker chart
    2. Visualization for composition
        (a) Pie and donut chart
        (b) Icon array

Bar Chart

- **Query:** What is the data split between “buy” and “did not buy” categories?
- This query will help us understand how the dataset is split between the two categories, which essentially is about checking the balance of the dataset.
- **Visualize:** A common way to visualize categorical variables is with a **Bar Chart** (Fig. 3.2). We can also use a Pie chart in this scenario as the number of categories is only two.

Bar Chart

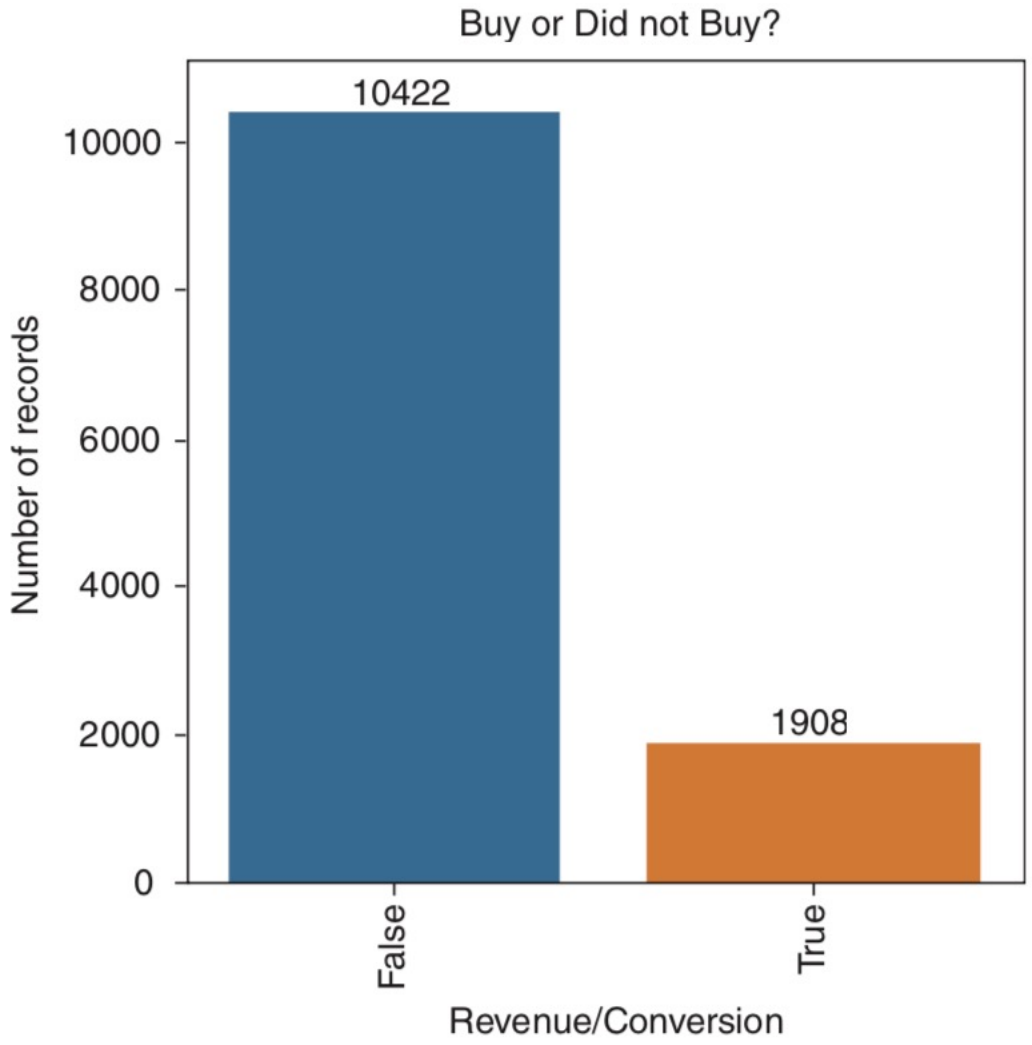


Figure 3.2 Bar chart- Visualize categorical feature.

- **Inference:** It is an imbalanced dataset as we only have 15.47% examples for “buy” (Fig. 3.2).
- The number of false classes is much higher than true classes.
- The classification model performance will suffer due to this imbalance in the data as most machine learning algorithms work best when the data is balanced.
- Hence, we need to consider sampling techniques such as up sampling and down sampling while we build ML models.

Big Number

- **Query:** What is the conversation rate for our e-commerce site?
- **Visualize:** When we want to convey an important message around a single number like conversion rate in this scenario, we can think of representing it as a Big Number.
- This is about using one huge number on our dashboard to catch the audience's attention.
- On the downside, care should be taken that it is used sparingly and not cluttered with too many of them on a dashboard.

Pie/ Donut Chart

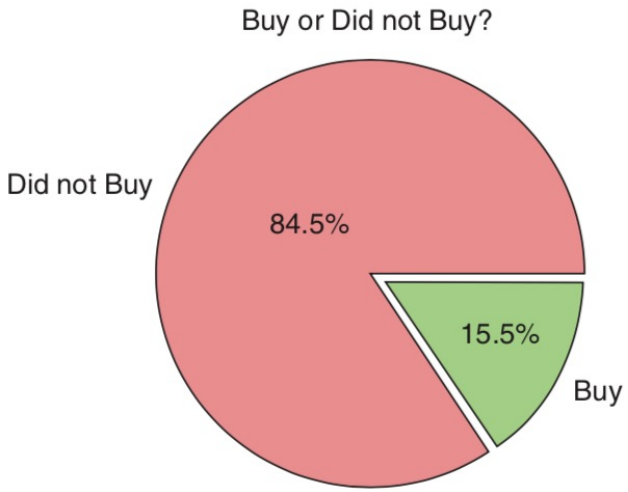


Figure 3.3 Pie chart.

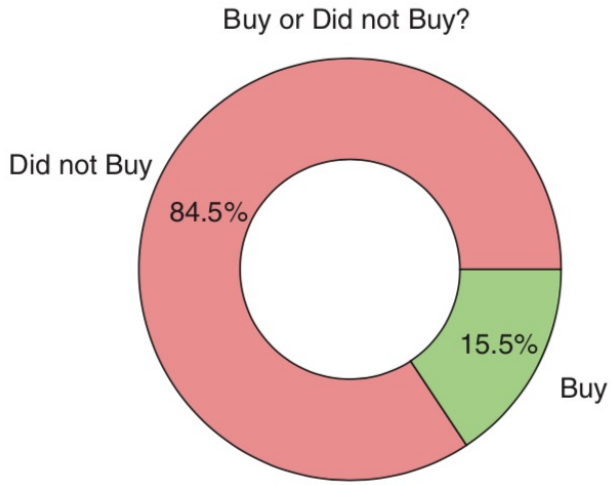


Figure 3.4 Donut chart.

- When we want to highlight a story around one part of a whole, we can use it by highlighting the piece of the pie as shown in Figs. 3.3 and 3.4.
- In general, estimating angles and curvatures is difficult for humans, hence pies and donuts are not ideal graphs.

Icon Array Chart

- We can utilize Icon array to tell a story around proportions.
- This would be useful to interpret if the audience has a lower ability to work with numbers.
- Figure 3.5 is Icon arrays generated using application available at http://www.iconarray.com. Icon array code is also freely available on GitHub.[2]

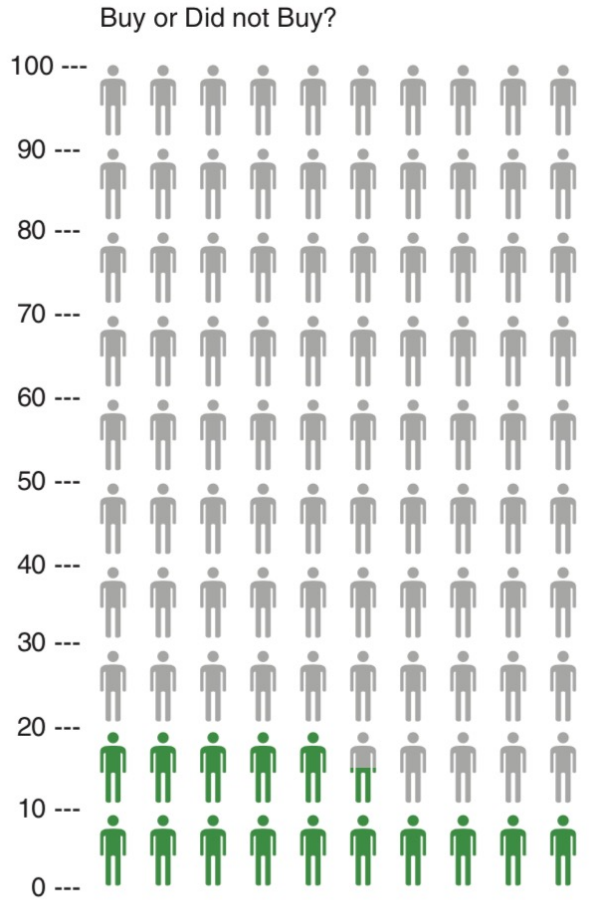


Figure 3.5 Icon array.

15.5 out of 100 people who visit us, buy from us

84.5 out of 100 people who visit us, Did not buy from us

[2]Images created by Iconarray.com. Risk Science Center and Center for Bioethics and Social Sciences in Medicine, University of Michigan. Accessed 2020-05-12.

Query: When do we get more shoppers on our website- week- days or weekend?

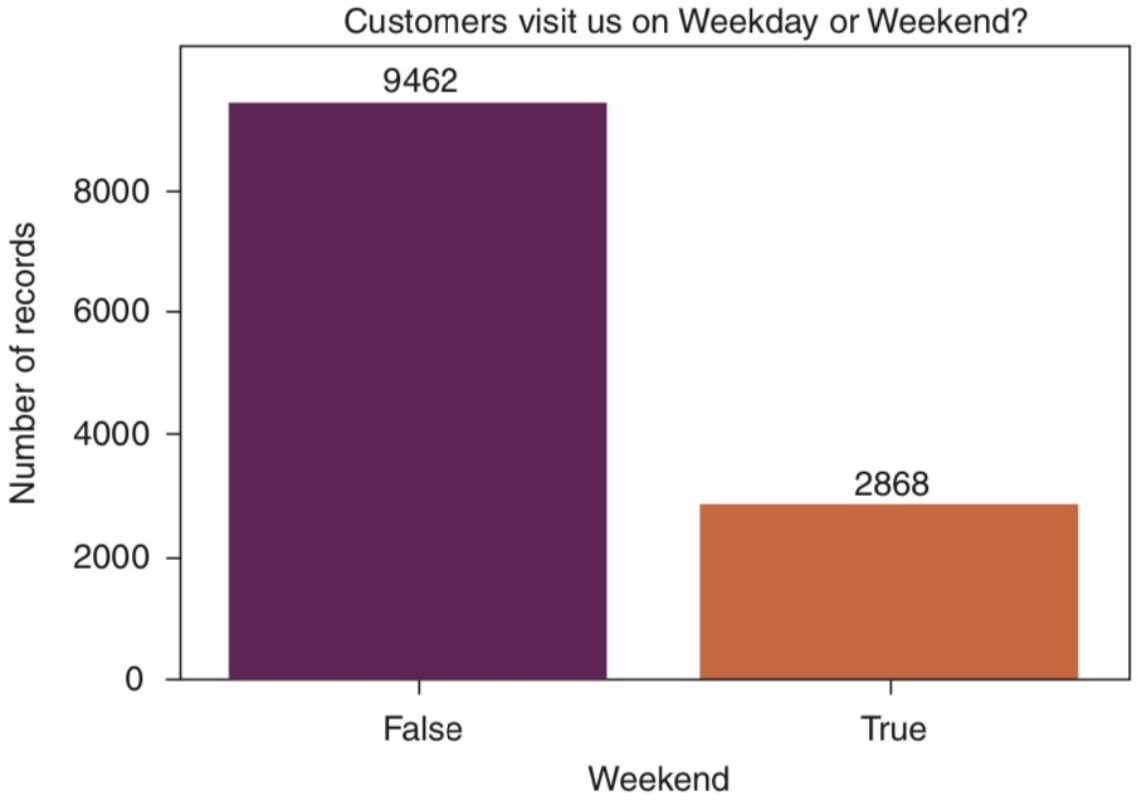


Figure 3.6 Bar chart on days customers visit.

- **Visualize:** Refer to Fig. 3.6.
- **Inference:** There are a large number of visitors during week-days rather than the weekend.
- One can further look into other factors in data to understand the possible reason behind it and gain insights for marketing department to run promotion campaigns during weekdays in order to maximize the reach.

Query: How is the break-up of visitors by visitor type?

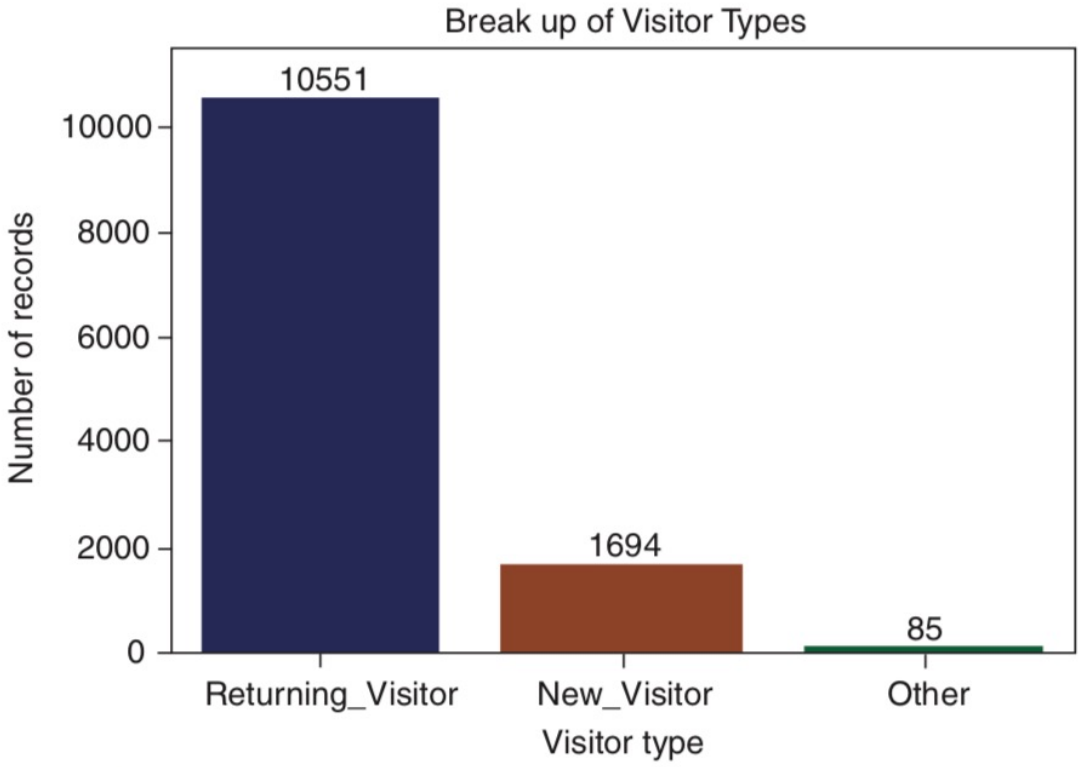


Figure 3.7 Bar chart on visitor types.

- **Visualize:** Refer to Fig. 3.7.
- **Inference:** We have the number of returning visitors significantly more than new visitors.
- This metric shows that the e-commerce is doing well in retaining its customers, as acquiring new customers is increasingly expensive.
- There are 85 records tagged under “other” visitor type.
- This needs to be checked with business on what type of visitors are tagged as others. These customers could be resellers, who buy in bulk from e-commerce sites and sell in retail.

Query: From which region do we get more traffic from?

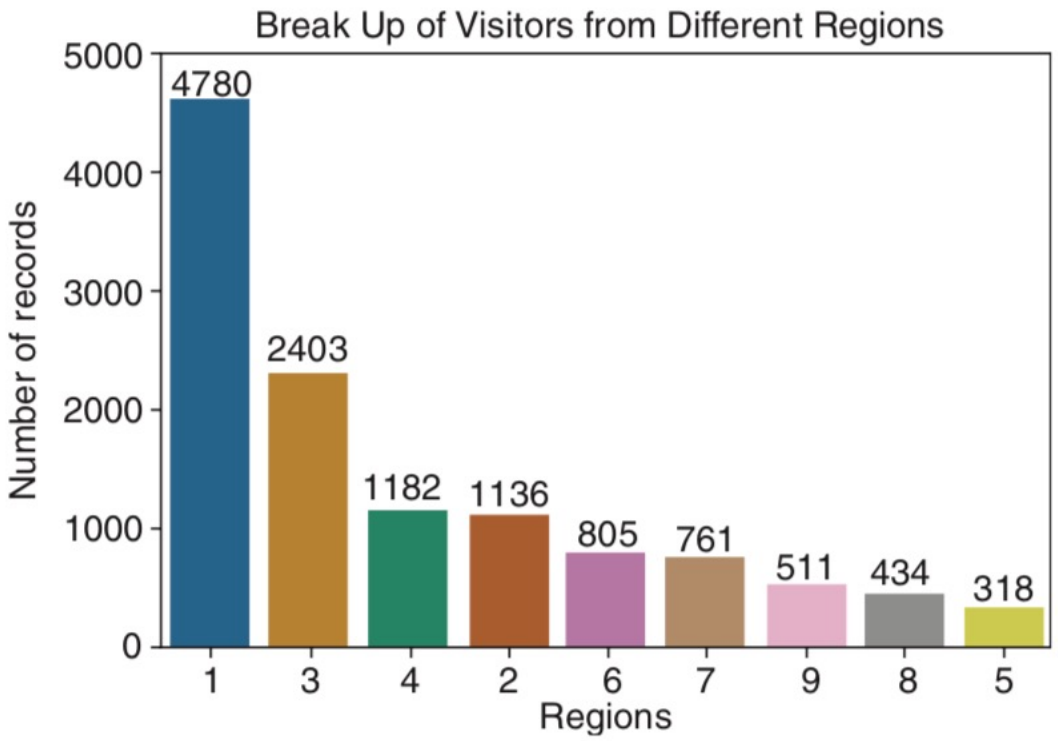


Figure 3.8 Bar chart on visitors from different regions.

- **Visualize:** Refer to Fig. 3.8.
- **Inference:** Largest traffic comes from region 1.
- Traffic from region 2 is lesser than that from regions 3 and 4.

Query: Is the source from where we get our traffic to website same across all source types?

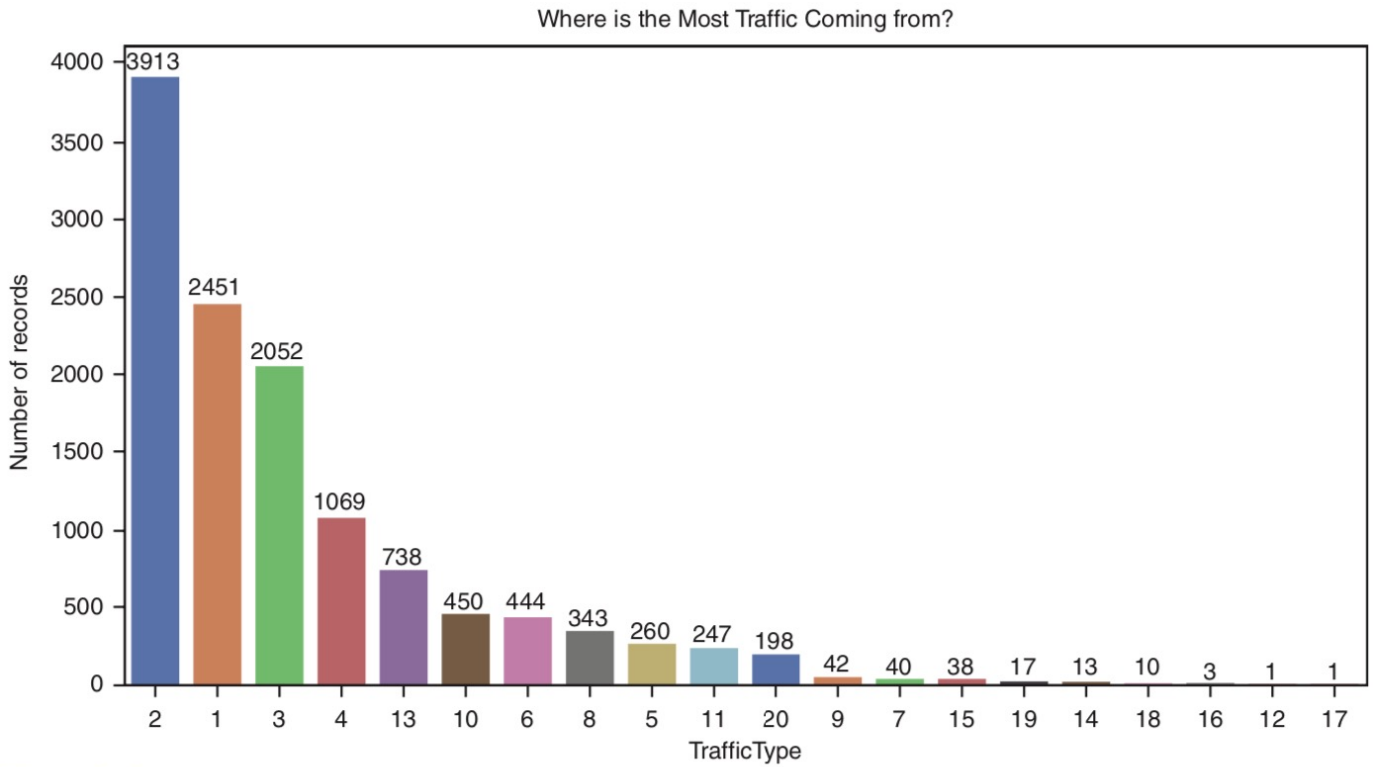


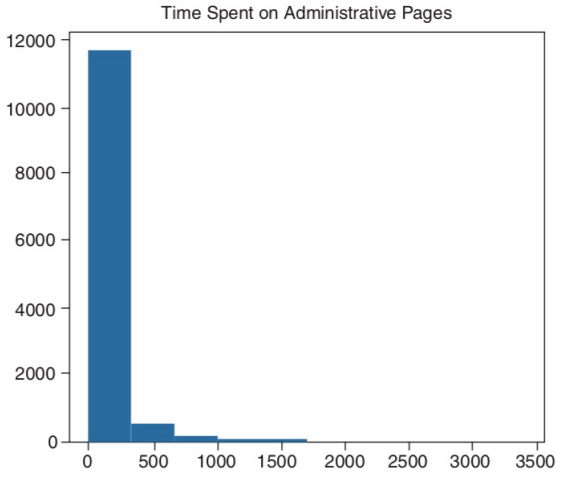
Figure 3.9 Bar chart on traffic types.

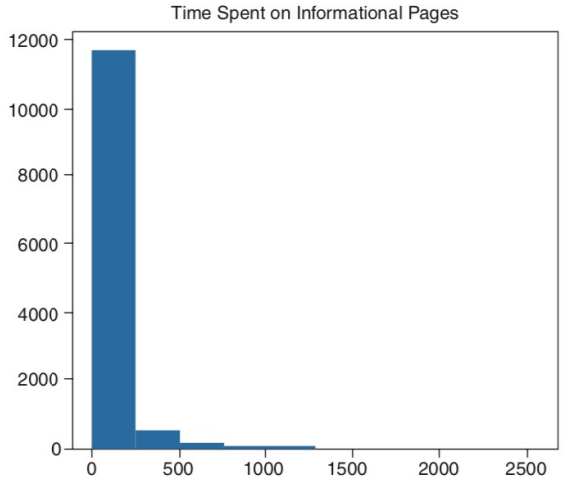
- **Visualize:** Refer to Fig. 3.9.
- **Inference:** Most of the traffic is routed from type 2, 1, 3, 4, and 13.

Histogram

- To analyse the variability in our numerical features, we use the Histogram chart.
- Histograms are a type of a bar chart where the vertical axis (*y*-axis) displays the relative frequencies or count of values in different class intervals.
- In histogram, instead of each bar representing a single value, it represents a range of values.
- It helps us to uncover skewness in the data.

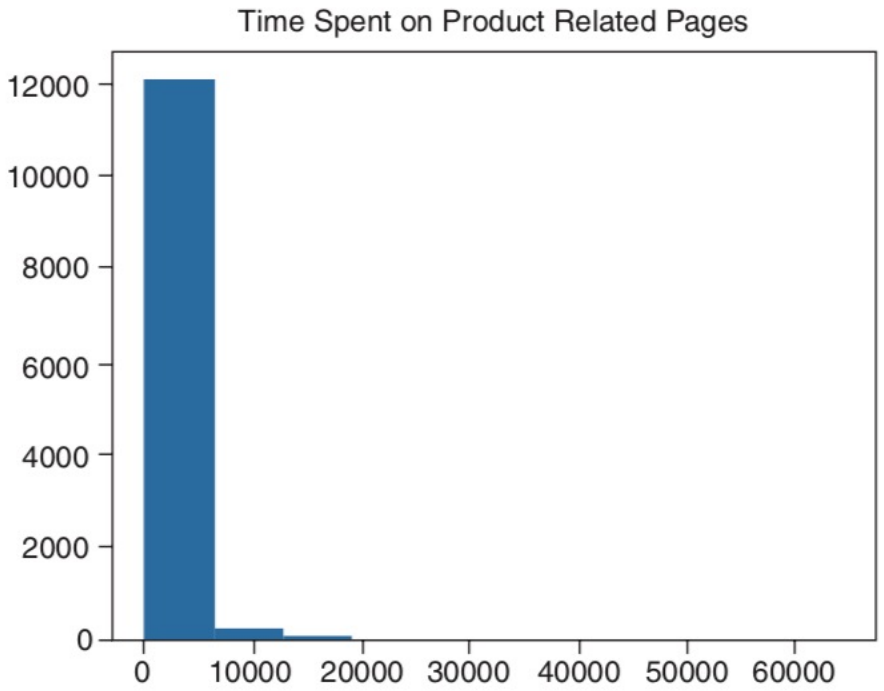
Query: Let us compare how much time visitors spend on the e-commerce site across pages, such as administrative, informational, and product-related pages.





- **Visualize:** Refer to Figures.
- **Inference:** The distribution of time spent across administrative, informational, and product-related page visits is highly skewed (right skewed).
- On administrative and informational pages, visitors spend maximum about 500 to 1000 seconds, while on the product related pages, visitors spend up to 15000 seconds.

Query: Let us compare how much time visitors spend on the e-commerce site across pages, such as administrative, informational, and product-related pages.



- **Inference:** Product-related pages are where product details and a call-to-action such as add-to-cart for e-commerce are located.
- These pages are also known as money pages because they most often invoke conversions.
- Hence, this indicates that visitors are more engaged by the content on product-related pages.

Query: What is the distribution of average number of pages that a user visits on an e-commerce website?.

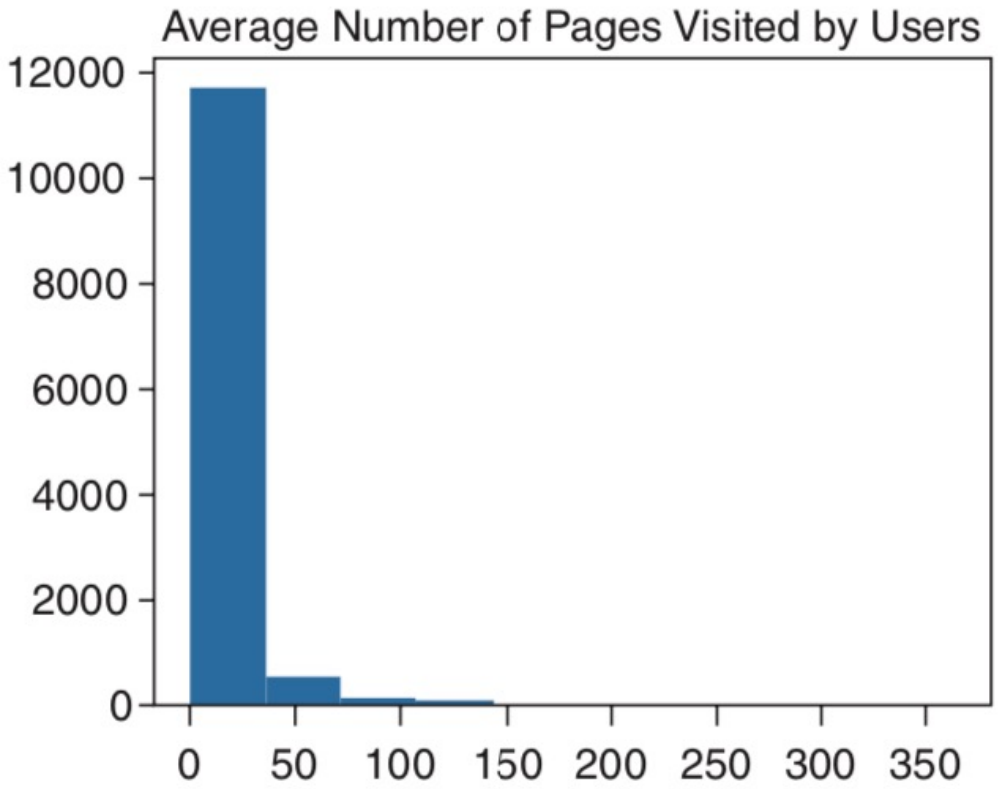


Figure 3.11 Histogram of average number of pages visited by customers.

- **Visualize:** Refer to Fig. 3.11.
- **Inference:** We can infer that most of the visitors browse between 0 and 150 pages before buying.
- According to a study done by the Nielsen Norman Group, half of purchase occurs within 28 minutes of the initial click (Nielsen, 2005).
- Hence, business/operations would strive for lesser page visits before purchase to ensure visitors are given compelling product details and quick action buttons to entice them to buy rather than loss of interest at customer level, which in turn corresponds to losing sales for the company.

Box and Whisker Chart

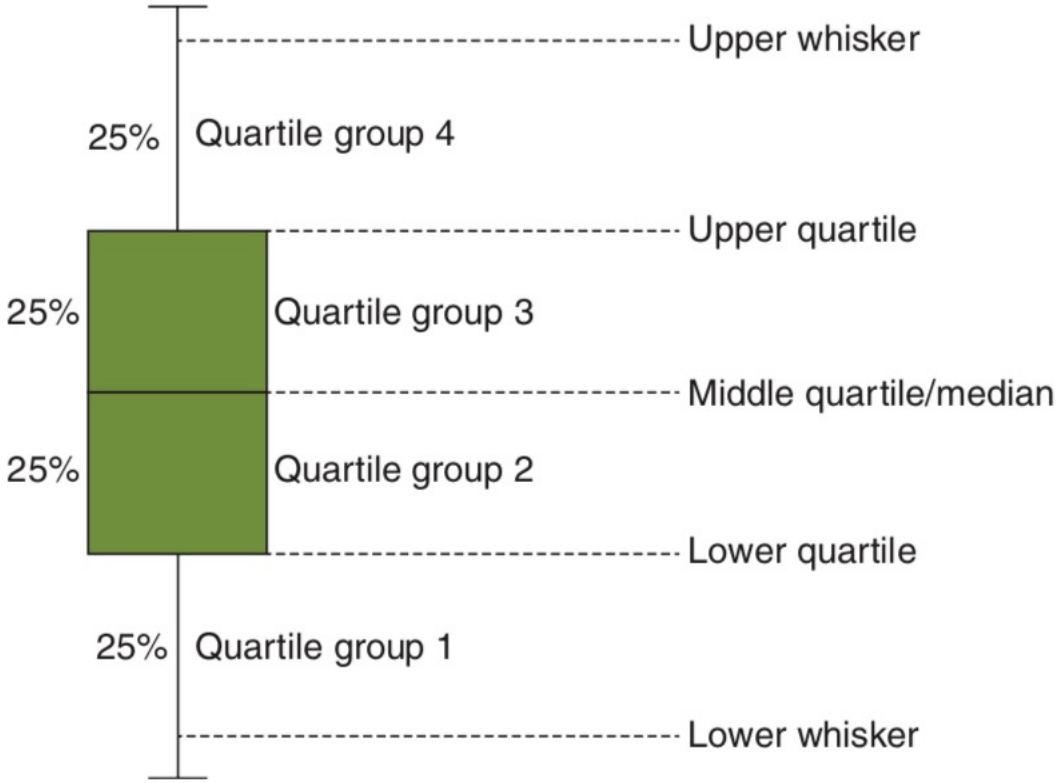


Figure 3.12 Box and Whisker Plot.
Source: https://wellbeingatschool.org.nz/sites/default/files/W@S_boxplot-labels.png

- A Box and Whisker Plot or Box Plot is used to visually display the distribution of data through their quartiles.
- The lower quartile (first quartile) represents 25% of the data has values lower than the lower quartile (group 1).
- The upper quartile (quartile 3) represents 75% of our data.
- The line within the box represents the median value of the data.

Box and Whisker Chart

- Inter-quartile range (IQR) is the difference between the lower and the upper quartile.
- The whiskers are defined as 1.5 times the IQR below group 1 and above group 3.
- Data points which fall outside the whiskers are considered outliers.
- The following observations can be made from viewing a Box Plot:
    1. **Key statistical measures such as the median, 25th percentile etc.**
    2. **Presence of outliers and their values**
    3. **Data symmetry**
    4. **Grouping of data**
    5. **Skewness in data**

Query: How is the distribution of visitor bounce rates of the e-commerce site?

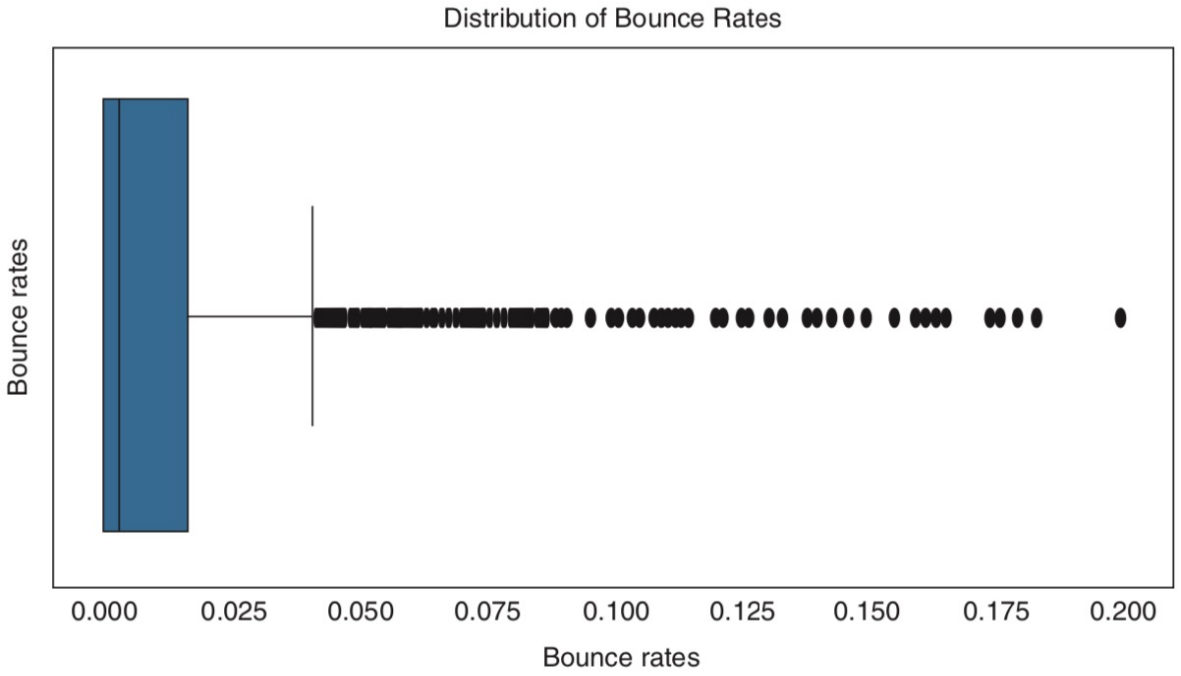


Figure 3.13 Box Plot to check distribution of bounce rates.

- **Visualize:** Refer to Fig. 3.13.
  Inference: We can infer that bounce rates data is positively skewed with the median at a very low bounce rate.
- Most of the bounce rates are between 0.000 and 0.050, but there are a few outliers (shown as dots in Fig. 3.13).
- Lower bounce rates would mean effectiveness of site pages for the business and operations teams.

Multivariate Analysis-Charts to Visualize Multiple Measure

- In Multivariate Analysis, we analyse multiple data features or attributes (two or more).
- Now we will create the following chart types to generate insights for the provided data:
  1. **Visualizing comparison**
     - **(a) Stacked bar chart**
     - **(b) Box chart and Boxen chart**
     - **(c) Violin chart**
     - **(d) Strip chart and Swarm chart**
  2. **Visualizing relationship**
     - **(a) Scatter chart and Joint plot**
     - **(b) Pair plot**
     - **(c) Heat map**
     - **(d) Parallel coordinates chart**
     - **(e) Dual axis chart**
  3. **Visualizing trends**
     - **(a) Line chart**

Stacked Bar Chart
Query: What is the conversion rate of the visitors

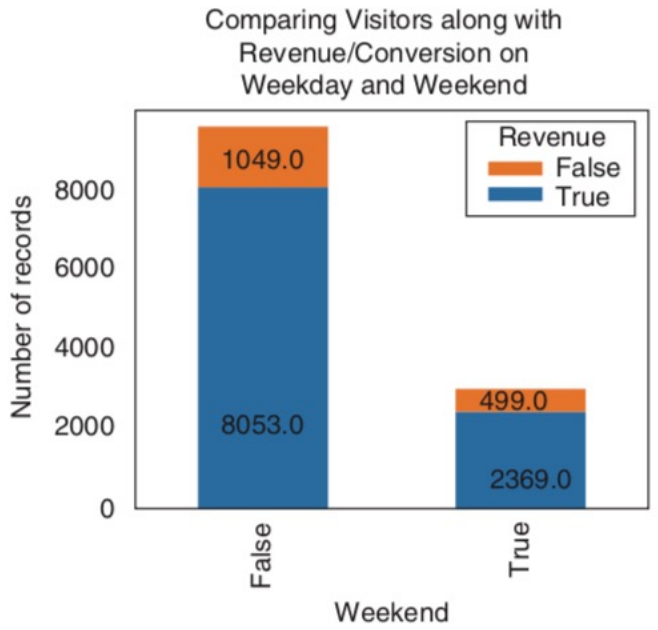


Figure 3.14 Stacked bar chart - Number of customers on weekday/weekend across revenue (buy or did not buy).

- **Visualize:** Refer to Figs. 3.14–3.16.
- In Fig. 3.6, we had “count of visitors” encoded as the length of the bars and “weekday/weekend” encoded as the colour of the bars and add the new dimension “revenue” and create a “stacked bar chart”.

Stacked Bar Chart
Query: What is the conversion rate of the visitors

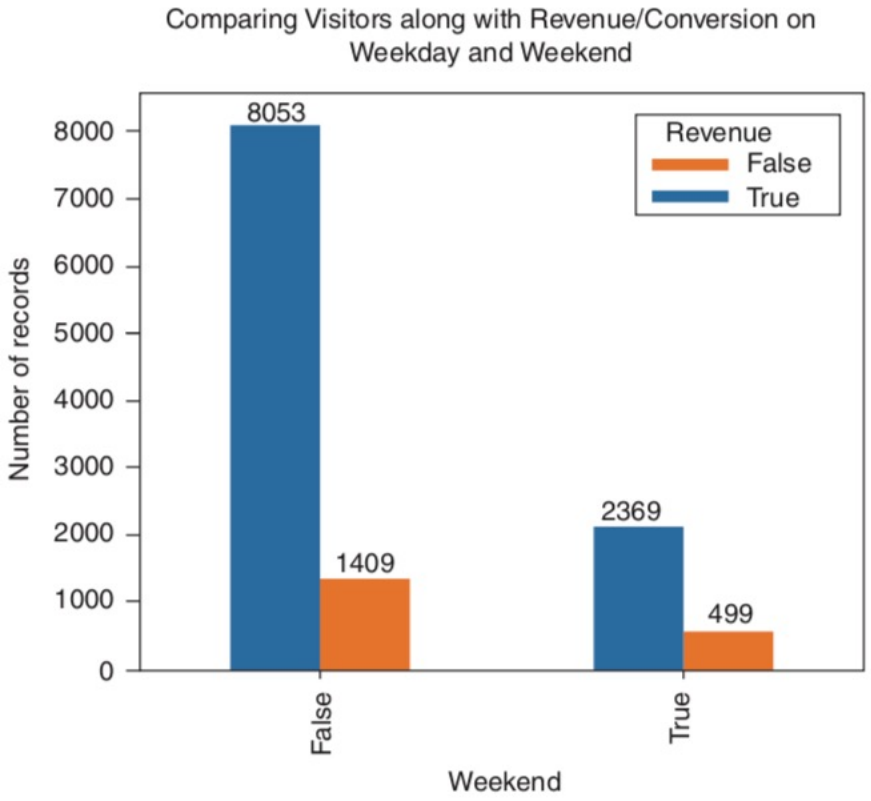


Figure 3.15 Grouped Bar Chart - Number of customers on weekday/weekend across revenue (buy or did not buy).

- We can either stack the "revenue" bars as in Fig. 3.14 or group the bars as in Fig. 3.15.

Stacked Bar Chart
Query: What is the conversion rate of the visitors

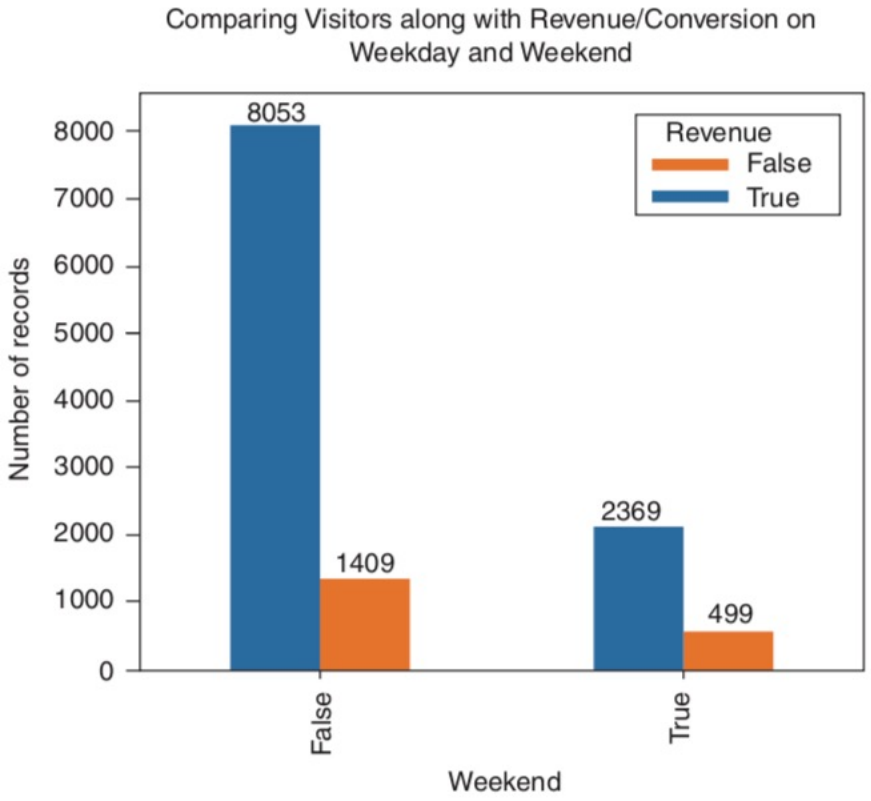


Figure 3.15 Grouped Bar Chart - Number of customers on weekday/weekend across revenue (buy or did not buy).

- We can either stack the "revenue" bars as in Fig. 3.14 or group the bars as in Fig. 3.15.

Stacked Bar Chart
Query: What is the conversion rate of the visitors

- The charts (Figs. 3.14 and 3.15) show that the number of visitors buying from the site during weekdays is more compared to weekends.
- However, if we look from a business/operations point of view, they would be more interested in comparing the “conversion rate”.
- But our chart still does not convey that information readily to the audience.

Query: What is the conversion rate of the visitors

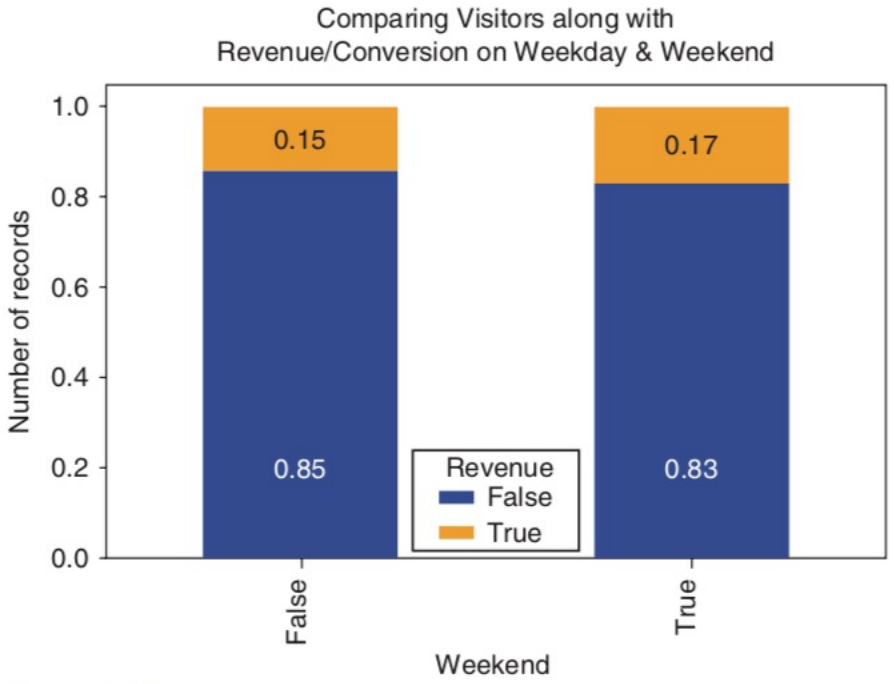


Figure 3.16 100% Stacked bar chart - % of Customer conversion across weekday/weekend.

- Hence, we can change our stacked bar chart as “100% stacked bar chart”
- Fig. 3.16, would enable us to compare the “conversion rate” for weekday and weekend.
- The only difference between 100% stacked charts and traditional stacked charts is that the bar or column in a 100% stacked chart is normalized to a consistent value of 100%.

Query: What is the conversion rate of the visitors

- It can be difficult to compare the proportion of each series to the whole if the total values of the bars/ columns greatly differ in a traditional stacked bar chart.
- But 100% stacked bar makes it easier to compare proportions within each bar or column by showing relative percentages.

- **Inference:** Even though more visitors come to the e-commerce site during weekdays, the revenue per visitor is marginally more over the weekends compared to weekdays.

Box Plot
Query: Compare the page values of the site with revenue

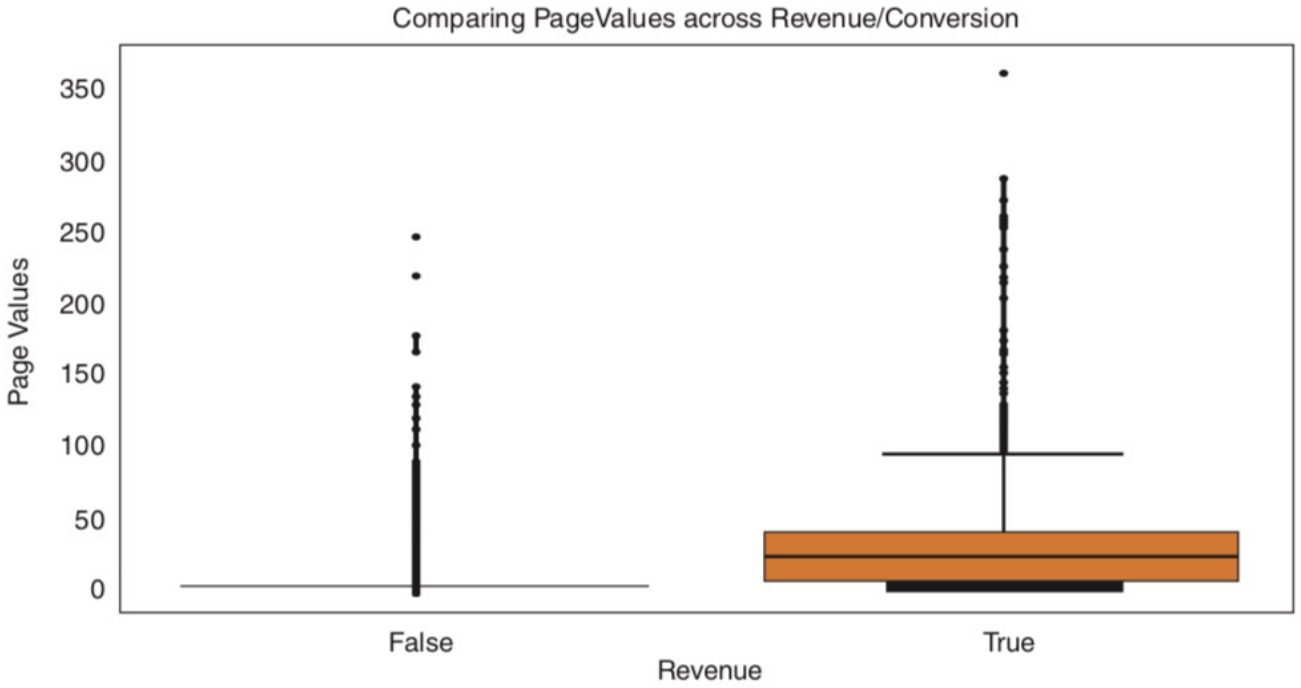


Figure 3.21 Box Plot - Compare average page views across revenue (buy or did not buy).

- **Visualize:** Refer to Fig. 3.21.
- **Inference:** Users are more likely to make the purchase when the page value is less than 50. Most of the users who did not buy are tightly grouped with page values close to 0.

Boxen Plot

Query: Does a visitor tend to buy when they spend more time on the administration page compared to Informational pages?

Boxen Plot

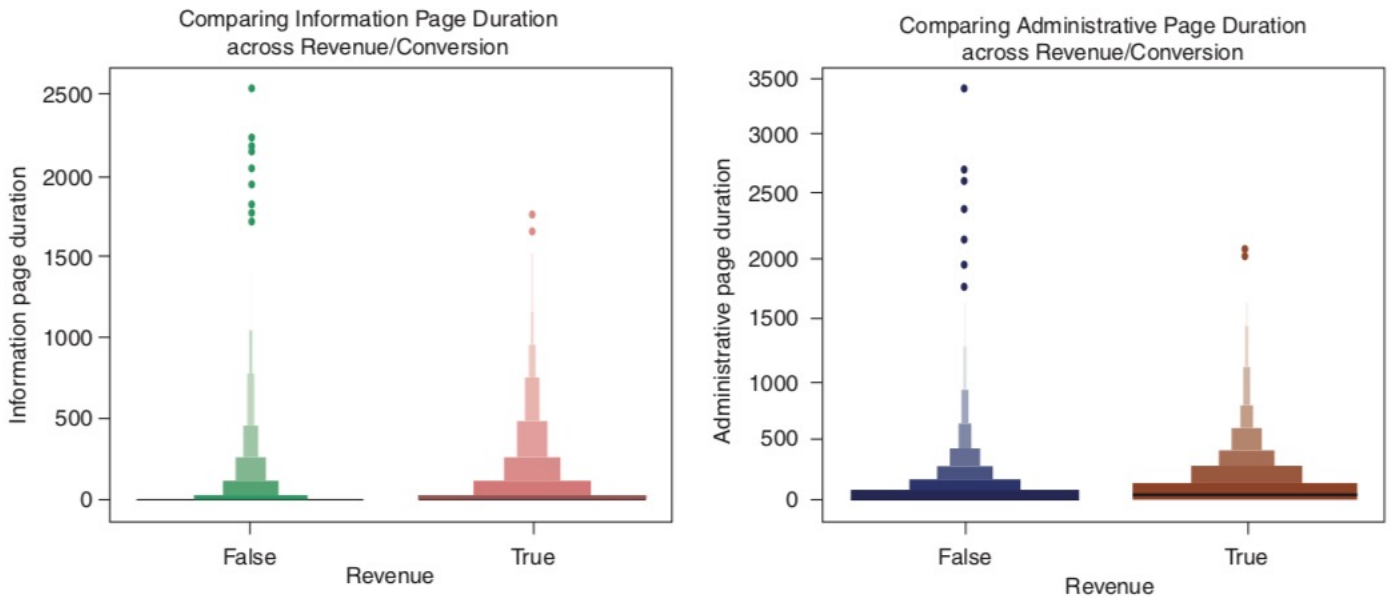


Figure 3.22 Boxen Plot- Compare time spent on informational pages and administration pages across revenue [true (buy) or false (did not buy)].

Visualize: Fig. 3.22

Inference: Customers visited administration pages more than informational pages. Also, customers visiting the administration page are more likely to make the purchase compared to those visiting informational pages.

Violin Plot

- The Violin Plot is like Box and Whisker Plots.
- We can plot the distribution of quantitative data in relation to the categorical variables, which in turn enables comparison between these distributions.
- In a Box Plot, the plot components are actual data points, but the Violin Plot features kernel density estimation of the distribution.
- Kernel density estimation is a statistical technique which helps us to create smooth curve of a given set of continuous data. It is way of estimating an unknown probability density function with given data (Rahul and U D Kumar, 2021).

Query: Does a visitor tend to buy when they spend more time on product related pages?

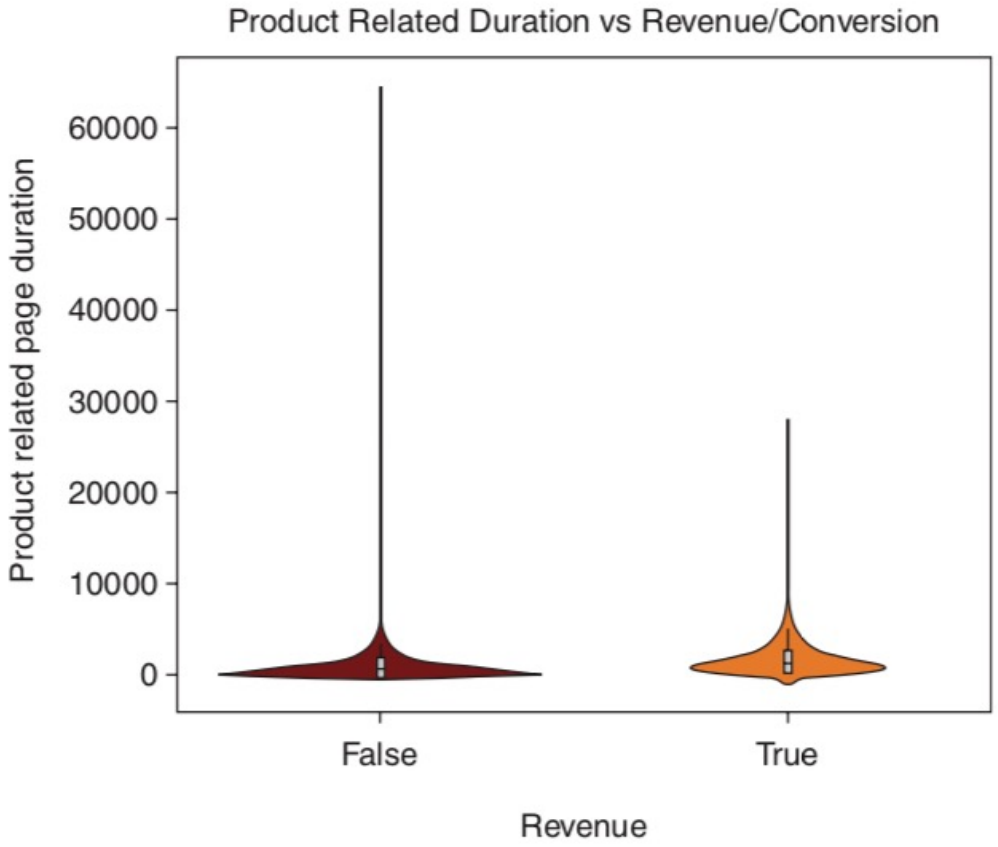


Figure 3.23 Violin Plot- Compare time spent on product-related pages across revenue (buy or did not buy).

- **Visualize:** Figure 3.23 shows the Violin Plot.
- **Inference:** Customers visiting product related pages are more likely to make a purchase compared to those who do not visit these pages.

Strip Plot
Query: Does a visitor tend to buy when they spend more time on product related pages?

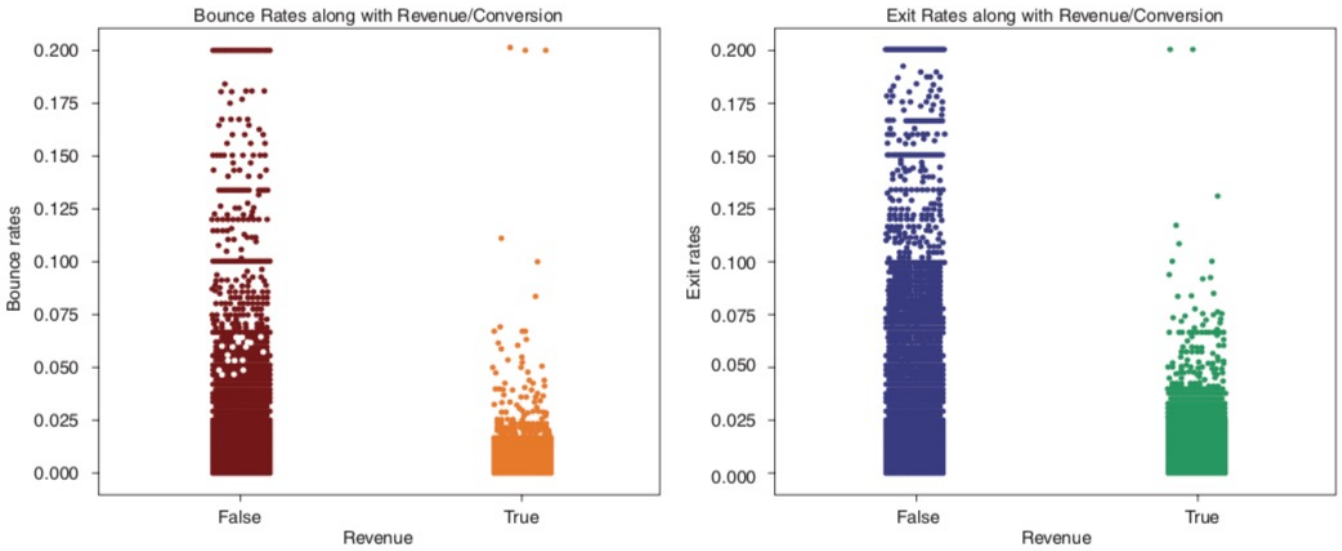


Figure 3.24 Strip Plot- Compare bounce rates and exit rates across revenue (buy or did not buy).

- A Strip Plot can be used to visualize all observations and some representation of what underlies those observations.
- **Visualize:** Figure 3.24
- **Inference:** Customers with low bounce rates and exit rates made purchases.

Swarm Plot

- Visually, a Strip Plot gets cluttered even with small data points (refer Fig. 3.24).
- If we want to have visual clarity of each and every data point, we can use the Swarm Plot.
- In Swarm Plot, each data point is spread out to avoid overlap. This provides a better visual overview of the data (Refer to Fig. 3.25).

Query: Compare visitor types, average Page values with revenue.

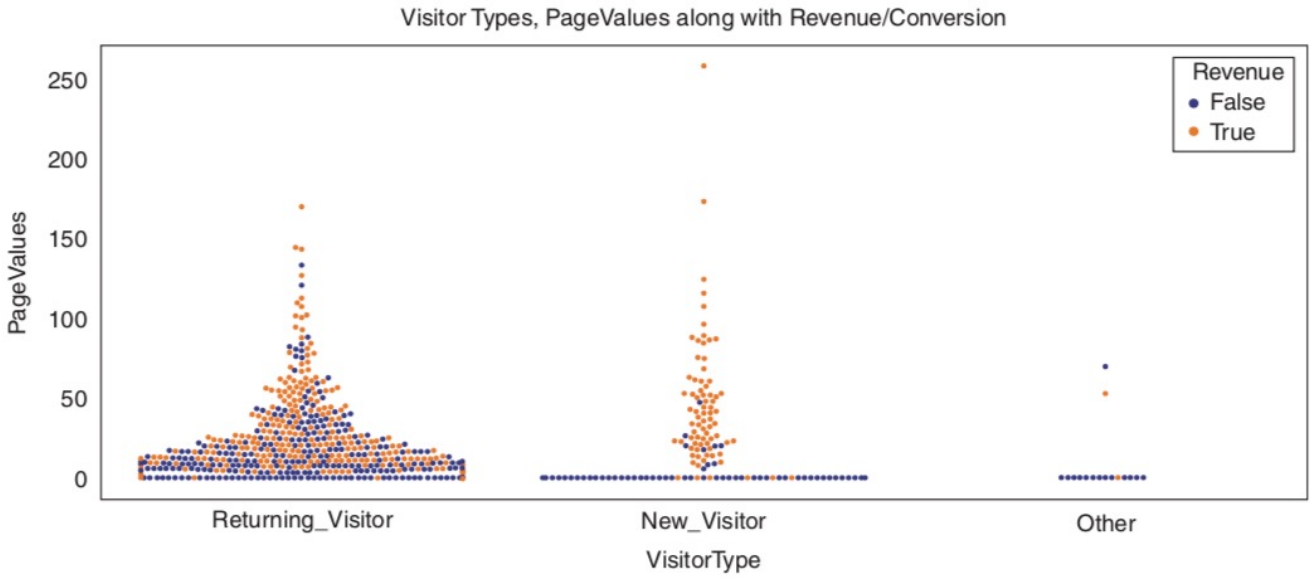


Figure 3.25 Swarm Plot- Compare visitor types, average page values across revenue [true (buy) or false (did not buy)].

- **Visualize:** Figure 3.25
- **Inference: T**here are lot of data points below 20 page views for a returning visitor as compared to new visitor.
- If we deduce that a returning visitor is aware of website layout as compared to a new visitor.
- Then from this swarm plot we can infer that before making a purchase, a new visitor browses multiple pages, which is not the case with a returning visitor.

Scatter Plot

- A Scatter Plot is primarily used to show relationships between two numeric variables.
- Data points on the horizontal and vertical axis show how much one variable is affected by another, which in turn helps us detect a relationship or correlation between the two variables.
- Relationships between variables can be linear or non-linear with positive or negative correlation (Refer to Fig. 3.26).
- It is also useful in identifying patterns in data by dividing data points into groups.
- Data points which are outside of the general cluster of points can be identified as outliers.

Scatter Plot

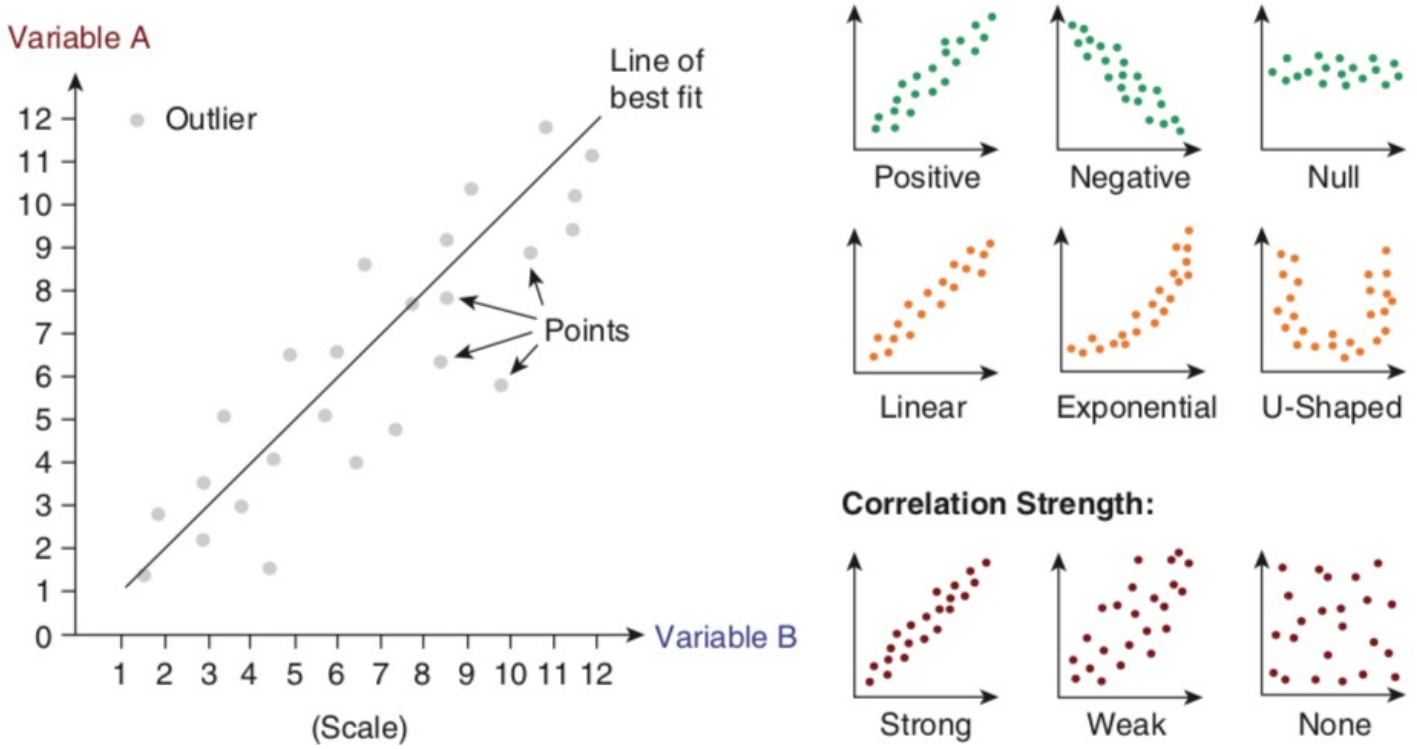


Figure 3.26 Scatter Plot.

Source: https://datavizcatalogue.com/methods/scatterplot.html

- Relationships between variables can be linear or non-linear with positive or negative correlation (Refer to Fig. 3.26).

Scatter Plot: Challenges

1. When we have a huge dataset, overplotting can be an issue with scatter plots. We can overcome this challenge by using sample data.
2. Care should be taken in interpreting Scatter Plot. There is a famous maxim, **“Correlation does not imply Causation”**. Hence, even though we see a relationship between two variables, it does not imply that changes in one variable are caused due to changes in the other.
    1. It is possible that a third variable (aka confounding or hidden variable) could be driving the changes in both variables.
    2. The third variable can be represented in a Scatter Plot by encoding it as colour, hue, shapes, and size.

Query: Is there any relationship between page values and Bounce and Exit rates?

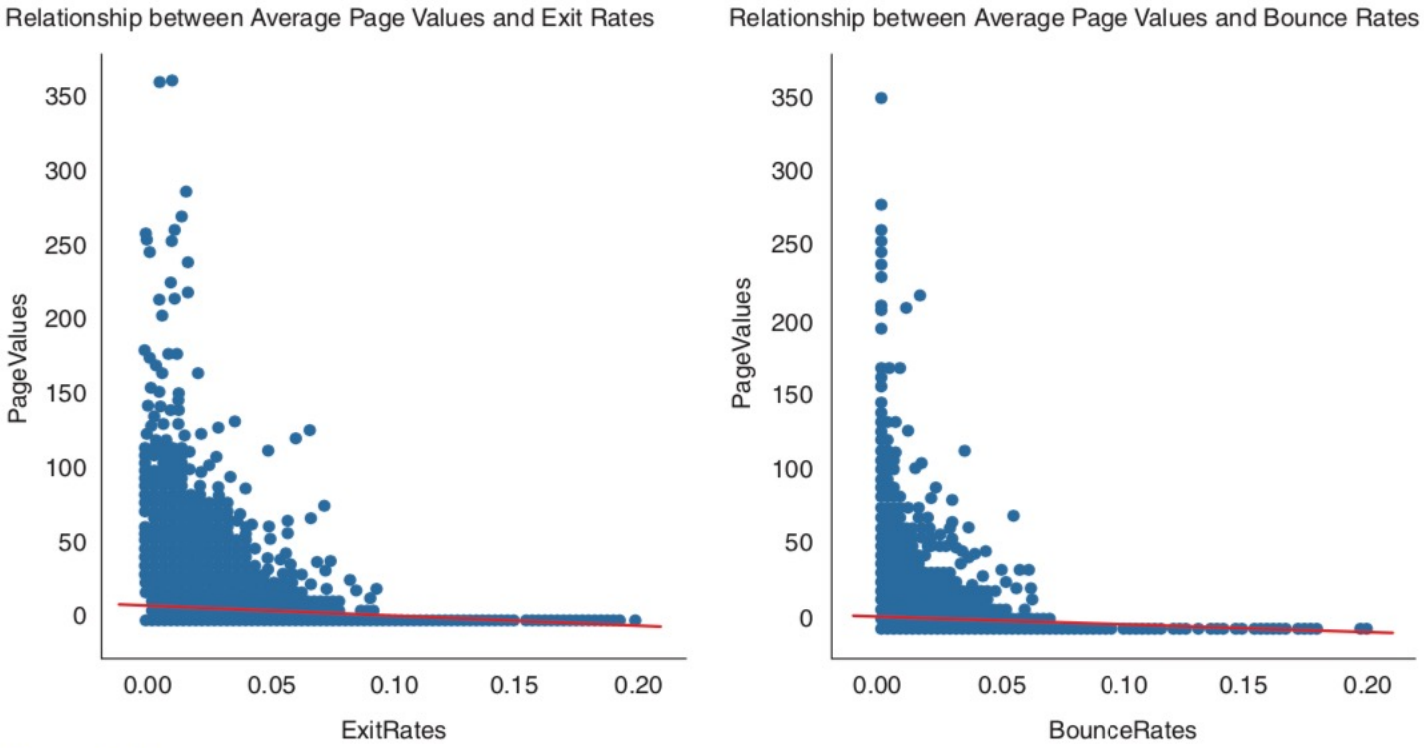


Figure 3.27 Scatter plot- Relationship between page values and bounce rates and exit rates.

- **Visualize:** Figure 3.27
- **Inference:** Higher bounce and exit rates are associated with lower average page values.

Joint Plot

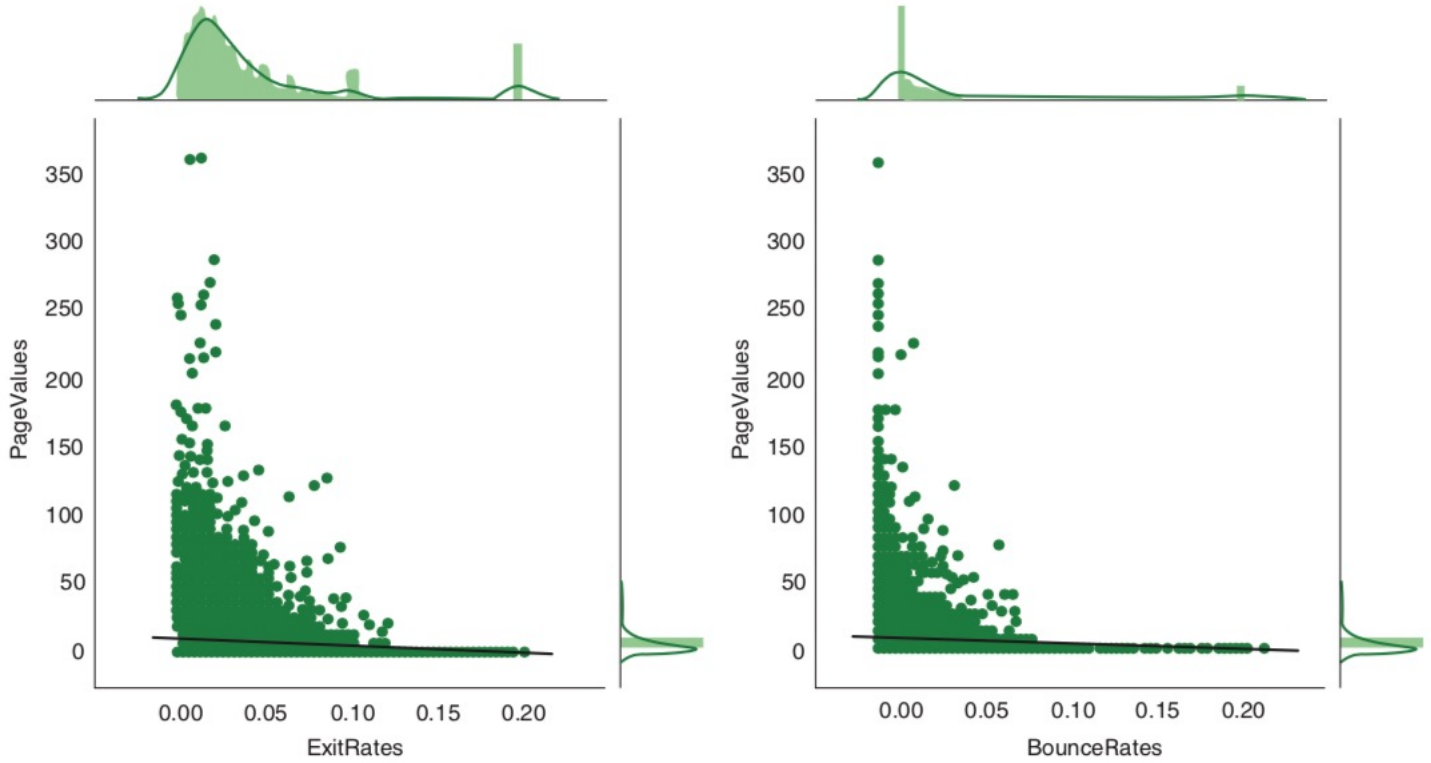


Figure 3.28 Joint Plot- Relationship along with distribution between page values and bounce rates and exit rates.

- The same details can also be visualized as a **Joint Plot**,
- In joint plot, we can check for correlations, relationships as well as individual distributions.

Pair Plot

- A Pair Plot helps us identify the relationship between a single variable and all other variables in the dataset.
- Pair plots help identify trends.

Query: Assess the relationship between some of our numerical features across revenue.

Pair Plot

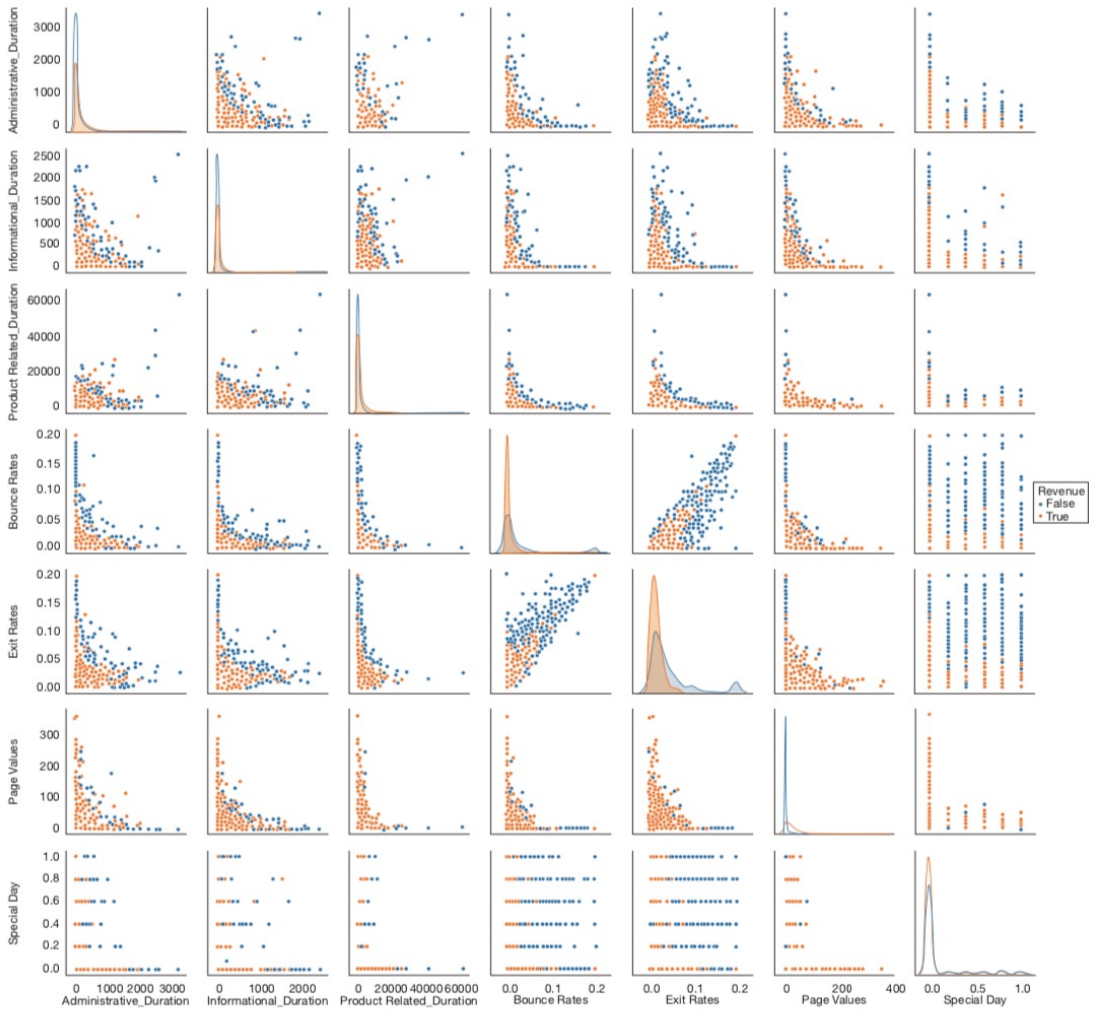


Figure 3.29 Pair Plot.

- **Visualize:** Refer to Fig. 3.29.
- **Inference:** Exit rate and bounce rate have a positive linear relation- ship. Customers with low bounce rates and low exit rates make purchases.
- A ‘special day’ does not have much effect on the customers’ purchase intention.

Heat Map

- A Heat Map is created from the correlation values of different features in the dataset.
- Correlation values are between −1 and 1.
- This plot helps us select features that correlate with others.
- In Fig. 3.30 values closer to 1 which indicates high positive correlation are all shaded in shades of orange, and values towards −1 which indicates high negative correlation are black in colour.

Query: Find a correlation between variables in the dataset.

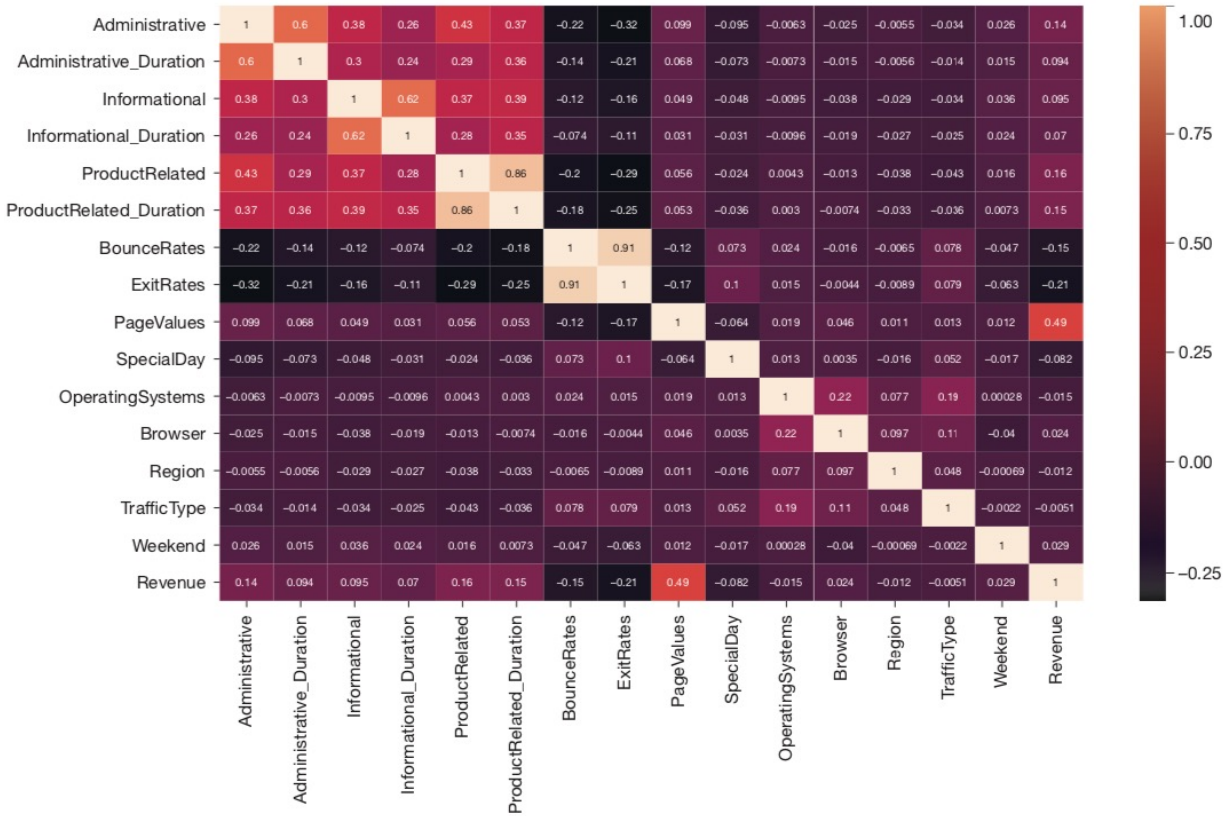


Figure 3.30 Correlation Heatmap.

- **Visualize:** Refer to Fig. 3.30.
- **Inference:** There is a correlation between average page values and revenue.
- As average page values increase, the possibility of a customer making the purchase also increases.
- Also, we see that people spending more time on product related page would have spent time on Informational and administration pages as well.

Parallel Coordinates Chart

- When we have multivariate data, Parallel Coordinates Chart is one way of visualizing multiple attributes together.
- It helps compare relationships among many variables in one chart.
- Each variable is represented as vertical axes parallel to each other.
- The data points are a series of connected lines across all the vertical axes.
- When creating a Parallel Coordinates Chart, care should be taken to arrange the vertical axes such that it enables the reader to understand the underlying data.
- Since it is easier to understand the relationship between the variables when the axes are adjacent to each other than when they are non-adjacent.
- We can try re-ordering axes to uncover patterns or correlations across all the variables.

Query: Evaluate the relationship between average page values, exit rates and bounce rates.

Figure 3.31 Parallel coordinates chart.

- **Visualize:** Refer to Fig. 3.31. Since average page values, exit rates and bounce rates are not on a common scale, we have normalized the values by scaling what aids in transforming raw data to a new scale that is common to all these variables.

Query: Evaluate the relationship between average page values, exit rates and bounce rates.

- **Inference:** In the visualization depicted in Fig. 3.31, line segments connect the data points.
- Each data attribute is represented as a vertical line.
- A single line segment connecting all the attributes represents one data point.
- Hence, points which are similar will appear closer together.
- We can infer from the visualization that most of our purchasing visitors having page values less than ~7.5–7.8.
- But average page values are spread across both high and low values for visitors who do not make a purchase.
- This kind of differentiation does not exist in Exit rates or bounce rates.

Line chart

- A line chart or line plot or line graph is a type of chart which displays information as a series of data points called ‘markers’ connected by straight line segments.
- In contrast to scatter plots, it consists of measurement points that are ordered and connected by straight lines.
- It is often used to visualize trend of one or multiple variables over time like analysing trends and patterns in financial data, comparing performance metrics of multiple groups, observing earnings per share of a company for different time cycles etc.

Query: Is there any trend in our bounce rates and exit rates over the last six months?

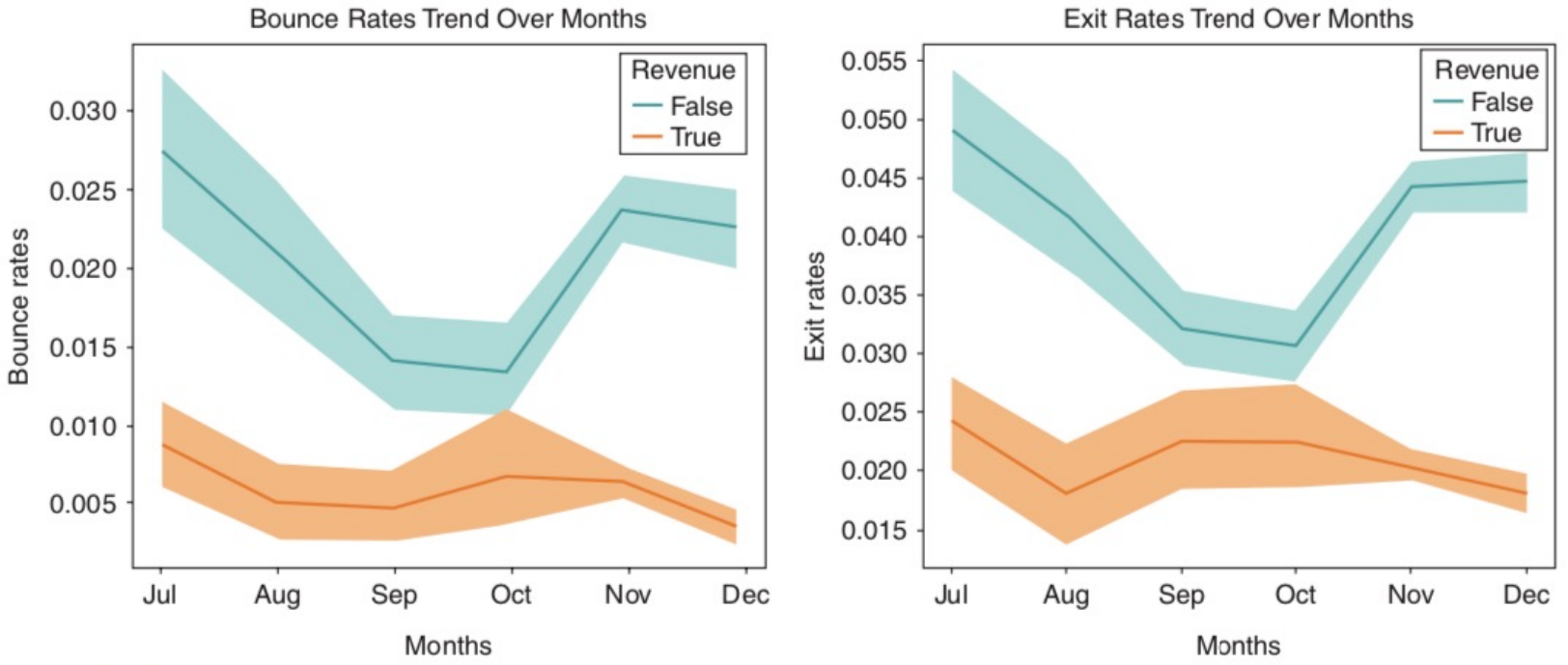


Figure 3.32 Line chart.

[4]https://en.wikipedia.org/wiki/Line_chart

- **Visualize:** Refer to Fig. 3.32.

Query: Is there any trend in our bounce rates and exit rates over the last six months?

- **Inference:** Bounce rates and exit rates trend broadly similar.
- The e-commerce site had higher bounce and exit rates during July, which improved till the end of September.
- Then, we see higher Bounce rates and Exit rates again in the last three months (October to December).
- This can be an indication to business/operations teams to check if there are any huge website changes in the last three months which might be triggering these rates.
- We also see a clear pattern- when bounce rates and exit rates are high, fewer customers convert a transaction into a purchase.

Dual Axis Chart

- The Dual Axis Chart or Multiple Axes Chart has two axes representing two variables with different magnitudes and scales of measurement.
- It helps us understand the relationship between two variables.

Query: Compare trends of revenue/conversion with number of visitors.

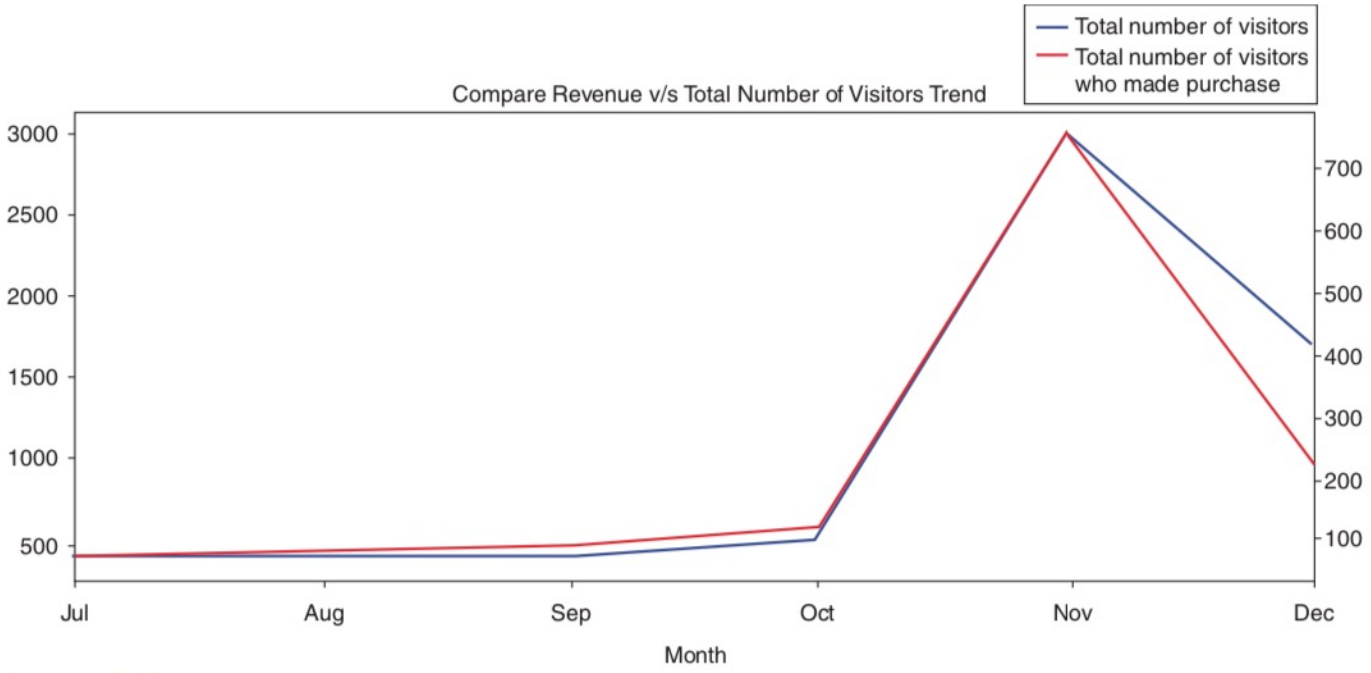


Figure 3.33 Dual Axis chart.

- **Visualize:** Figure 3.33 shows double axis chart.
- **Inference:** We can infer that as we have more customers/visitors on the e-commerce site, the revenue/conversion also increases.
- This is an insight for business/operations to devise marketing strategies to draw more visitors to e-commerce sites.

Modelling

- In this section, we will discuss how data visualization is used for analysis in key stages of the machine learning model development process.
- Data visualization is utilized for
    1. analysing key features,
    2. selecting the best model, and
    3. for tuning hyper-parameters.

Feature Selection

- Feature Selection is an important step for successfully building robust ML models.
- Feature Selection allows data scientists to select a feature of importance in predictive analytics.
- The sample dataset discussed in this chapter is taken from the UCI repository,
- The goal of a feature selection is to find the smallest set of available features which will help us predict the outcome at maximum accuracy.

Feature Selection: Why do we have to reduce the number of features?

1. Minimizing the number of features will help us lower the complexity of the model, which in turn reduces model bias.
2. It takes less computational time for modelling and prediction with Lower dimensional data.
3. It is easy to interpret a model built using a fewer number of variables.
4. Maintaining the model is less expensive.

Feature Selection

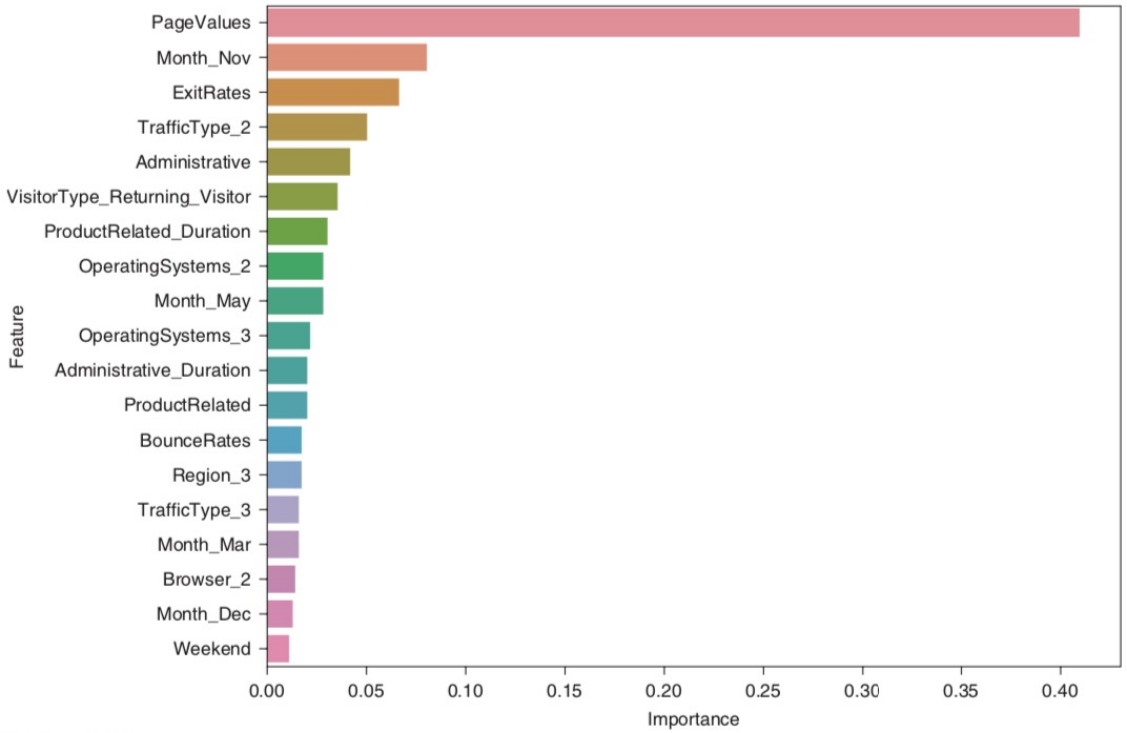


Figure 3.34 Feature importance chart.

- Feature importance of each feature is found using the feature importance property of the model.
- For every feature in the data, feature importance provides a score.
- Features with high importance and relevance towards the output variables will have a higher score.
- A random forest model was used to check for important features.

Hyper-parameter Tuning

- Hyper-parameters are the parameters whose value must be decided before machine learning model building.
- Once we have selected features to be used to build our model, we need to tune our model with the best parameters.
- We plot and check the model's training and validation scores for a single hyper-parameter.
- Random forest model has hyper parameters such as number of trees, number of features used in model development and tree depth.
- In our Radom Forest classifier model, let us visualize what the ROC scores are when we change our hyper- parameter—Tree Depth.

Hyper-parameter Tuning

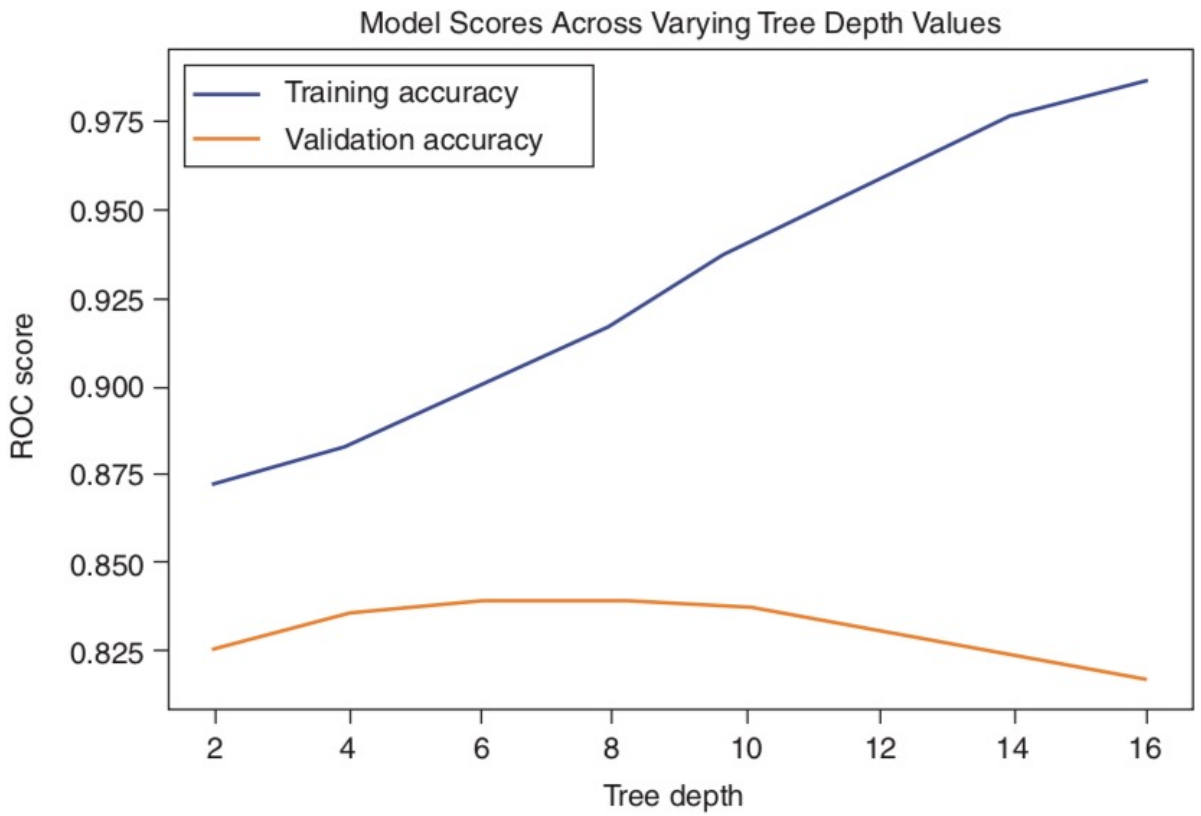


Figure 3.35 Hyper-parameter Tuning.

- As we can see from Fig. 3.35, Tree Depth = 8 seems to be the value where the ROC Score is high, both in the training and the validation data.
- ROC can be used to understand the overall worth of the classification model (Kumar, 2021).

Model Evaluation

- To build better intuition around the performance of the model, data scientists use visualizations along with the numeric scores.
- In the random forest classifier, which we have built for prediction of “customer intention to buy or not” data.
- We are interested in analysing values predicted by the model compared to the actual labelled values of the dataset.

Confusion Matrix

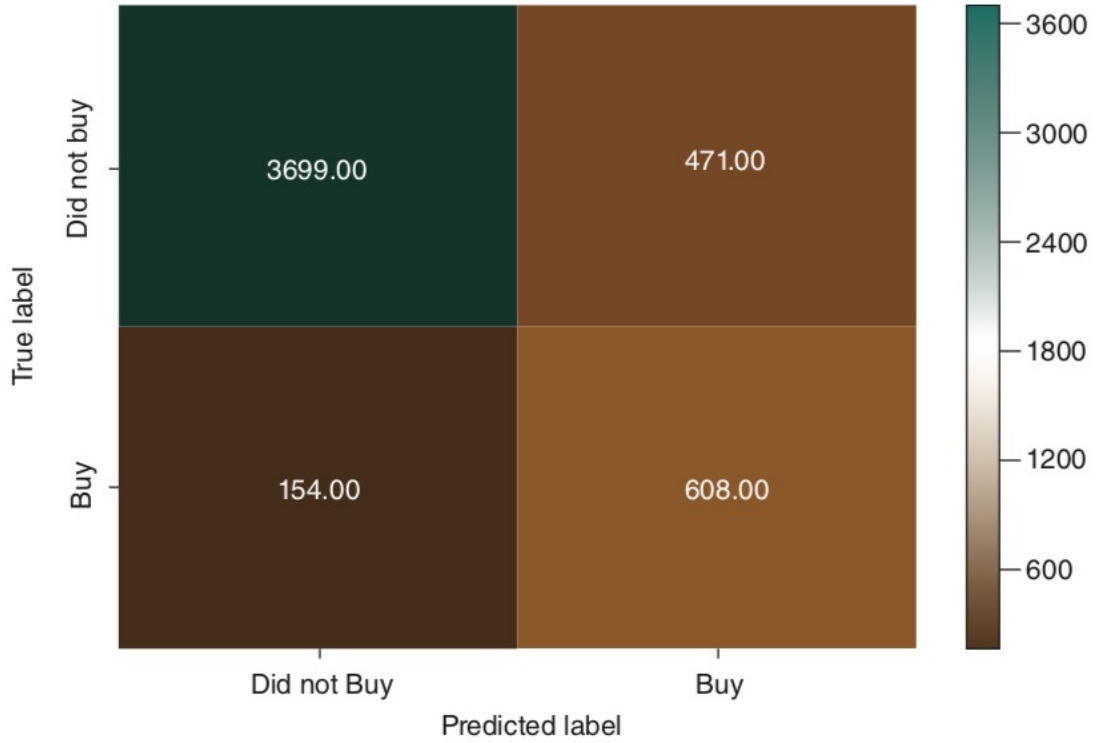


Figure 3.37 Confusion Matrix.

- In this prediction,
  1. We would be right:
     When we can classify the customers who buy as “buy” (True positives) and customers who did not buy as “did not buy” (True negatives)
  2. We would be wrong:
     When we classify customers who buy as “did not buy” (False negatives) and customers who did not buy as “buy” (False positives)

Classification Report

- Confusion matrix can sometimes be difficult to understand, particularly when we have more classes to predict.
- It only depicts the number of instances which have been classified properly or improperly.
- Hence, we can use a classification report which provides the data on three evaluation metrics: precision, recall, and F-Score.
- The classification model performance is measured using these metrics.

Classification Report

- Precision is the conditional probability that the actual value is positive, given that the prediction by the model is positive.
  - Precision = *TP / (TP + FP)*

- Recall, also referred to as sensitivity, is the conditional probability that the predicted class is positive given that the actual class is positive.
  - Recall (Sensitivity) = *TP /(TP + FN)*

- F-Score is a measure that combines precision and recall (harmonic mean between precision and recall).
  - F-Score = (2 × Recall × Precision)/ (Recall + Precision)

Classification Report

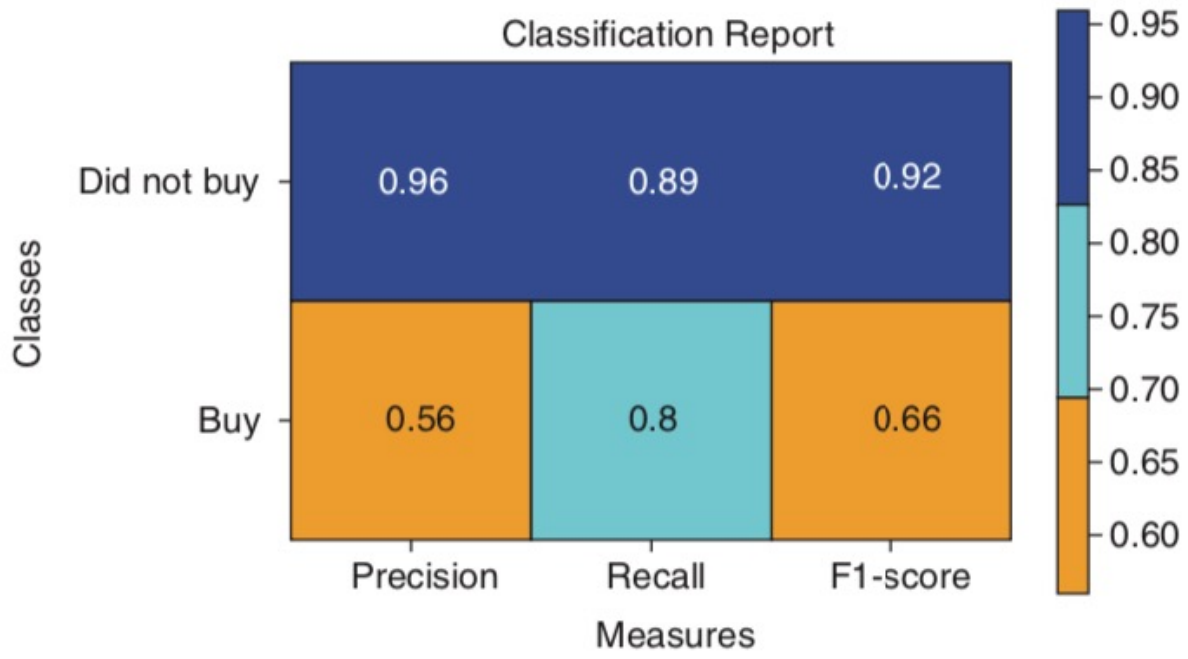


Figure 3.38 Confusion Matrix.

- A classification report can be visualized as a color-coded heat map.
- Fig. 3.38 helps us understand the model metrics in terms of successful prediction (blue colour) and weak predictions (orange colour).

ROC_AUC Plot

- We can use Receiver Operating Characteristic (ROC) and area under the Receiver Operating Curve (AUC) plot to examine the performance of the model.
- ROC_AUC plot helps us visualize the trade-off between the model's sensitivity and specificity.
- Sensitivity is how well the model has optimized to find true positives, and
- Specificity is how well the model has optimized to avoid false positives.

ROC_AUC Plot

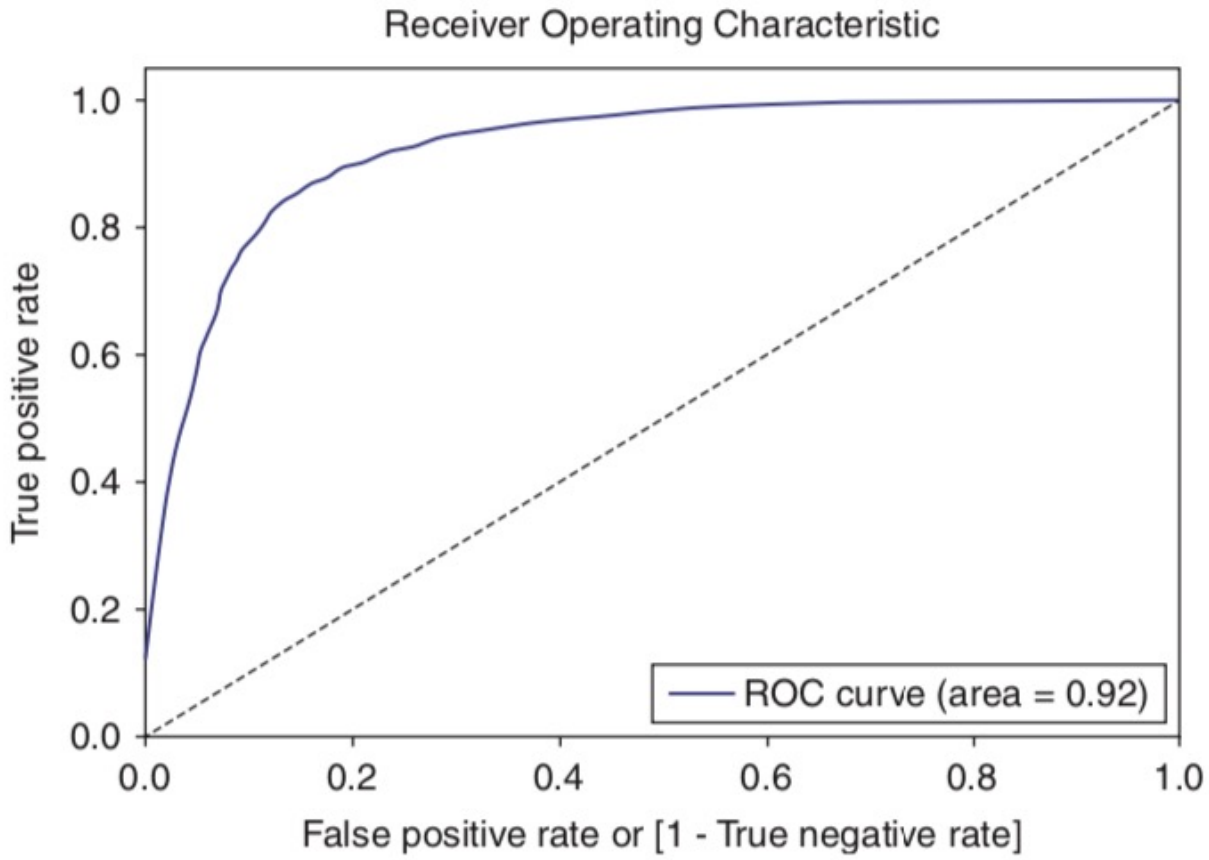


Figure 3.39 ROC_AUC plot.

- In Fig. 3.39, the horizontal axis is the False Positive Rate (1 - Specificity) and the vertical axis is the True Positive Rate (Sensitivity).
- In the plot, the model will have good accuracy when the curve is pulled toward the upper left corner.
- The model is as effective as a random coin toss when the curve is aligned with the diagonal line.

Visualization during Deployment

- In any data science project, we will have multiple skillsets in our audience who will be using the developed model.

1. Technical team who will maintain the model: These are the technical/functional audience who will be interested in under- standing how the model makes decisions.
2. Business/Operations teams who make decisions/build strategy: These are audiences with functional domain expertise who will be interested in knowing how to take decisions using the model.

Visualize Decision Trees

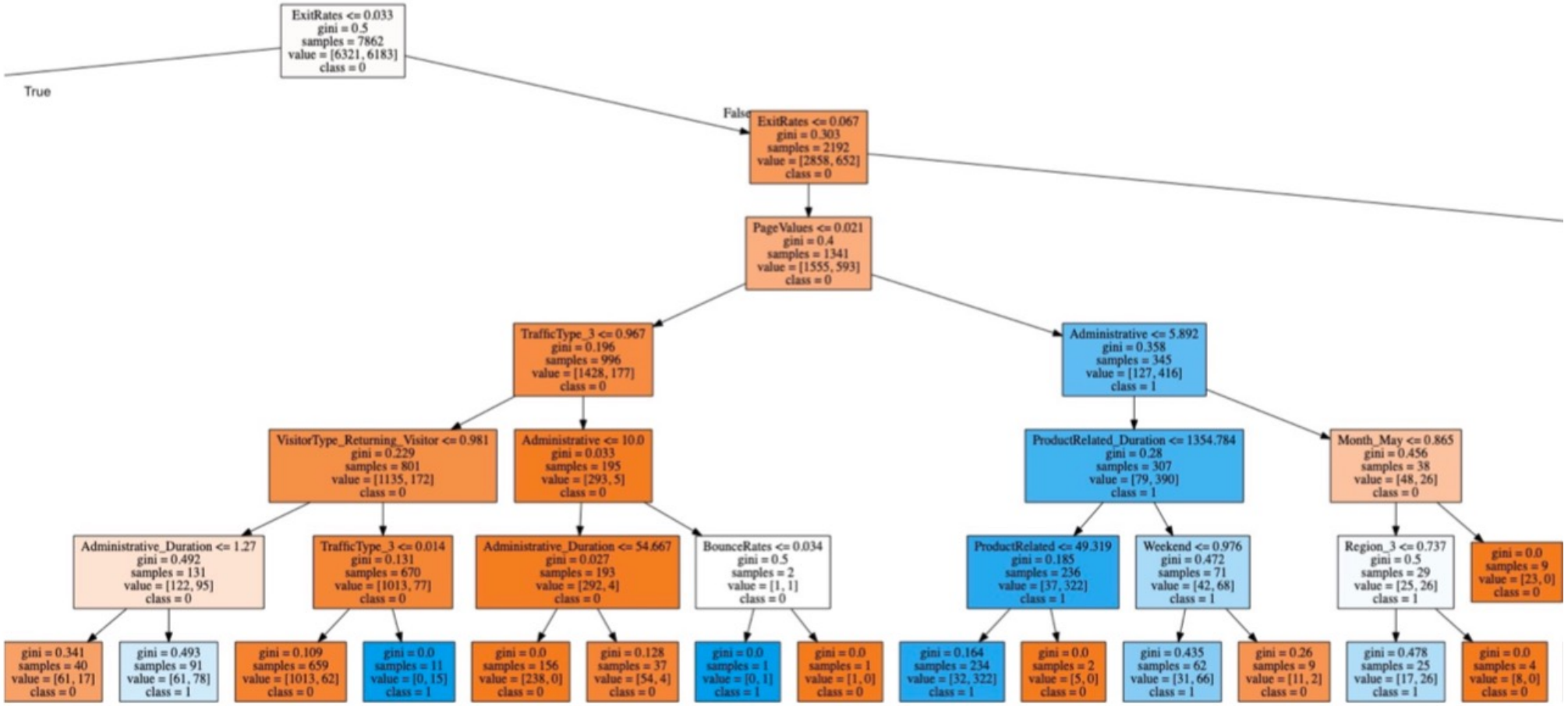


Figure 3.40 Partial decision tree.

- Fig. 3.40, visualize one of the trees used in the random forest model.

Model Explainer

- Explainer is a good approximation of the complex ML model. Using model explainer, we can interpret the complex model as:

1. **Local interpretation** – This would be for a specific prediction. This will help in understanding the linear combination of fea- tures used to predict the specific observation.
2. **Global Interpretation** – This would be used to explain how our model works overall.

Local Interpretation

- Consider an observation from our test dataset which had actual outcome as “buy” and our random forest model also predicted “buy”.
- Using local Interpretation visualization, we can understand features which aid the prediction as “buy”.
- We can also visualize the features which are moving this prediction towards “no buy”.

Local Interpretation

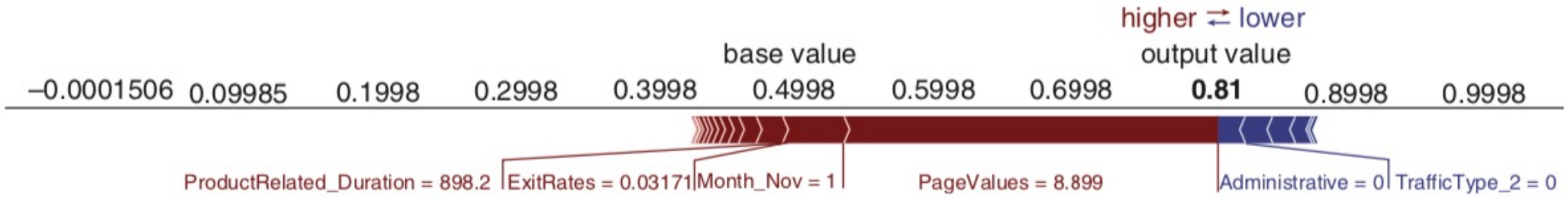


Figure 3.41 SHAP local Interpretation.

1. Base value = 0.4998. This is the expected value of the model; that is, what our model would have predicted without any information.
2. Output value = 0.81. This is what our random forest model predicted as the probability of buy for this observation.
3. Features with SHAP values – SHAP values is an acronym from SHapley Additive exPlanations. It breaks down a prediction to show the impact of each feature. SHAP values of all features sum up to explain why the prediction was different from the baseline.

Local Interpretation

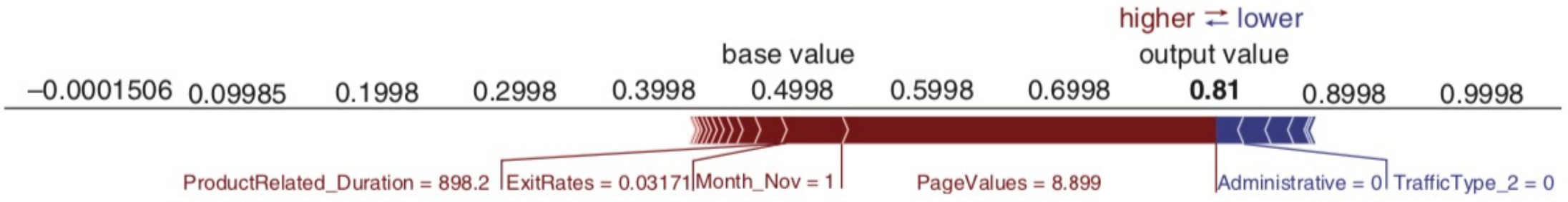


Figure 3.41 SHAP local Interpretation.

- We can see there is a shift from the base value towards output, and the features (marked in red arrow towards right) which are aiding this shift.
- It says page values feature contributed largely, followed by exit rates and administrative features.
- These features increase the probability of our correct prediction.
- These are also called positive SHAP values, features which contribute positively to the outcome.

Local Interpretation

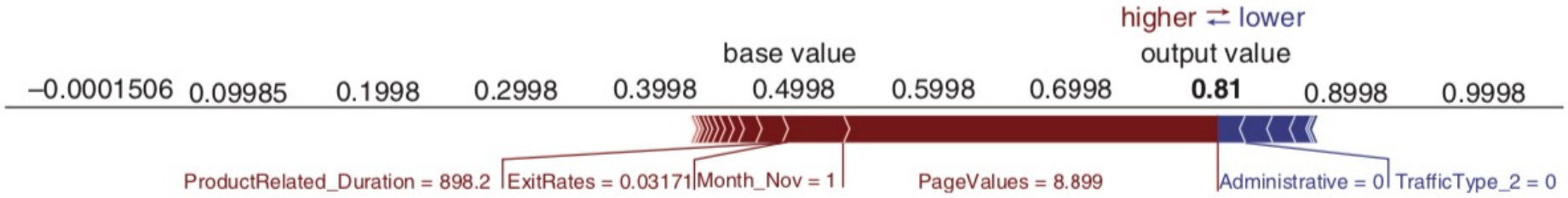


Figure 3.41 SHAP local Interpretation.

- Features (marked in blue arrow towards left) which try to decrease the probability such as month, traffic type and visitor type.
- These are also called negative SHAP values, features which contribute negatively to the outcome.
- For this observation,
    1. Page values = 8.899, Month = November and exit rates = 0.03171 have positive impact, increasing the probability of the customer intention to buy.
    2. Administrative=0andTraffictype_2=0havenegativeimpactin decreasing the probability of this customer intention to buy.

Global Interpretation

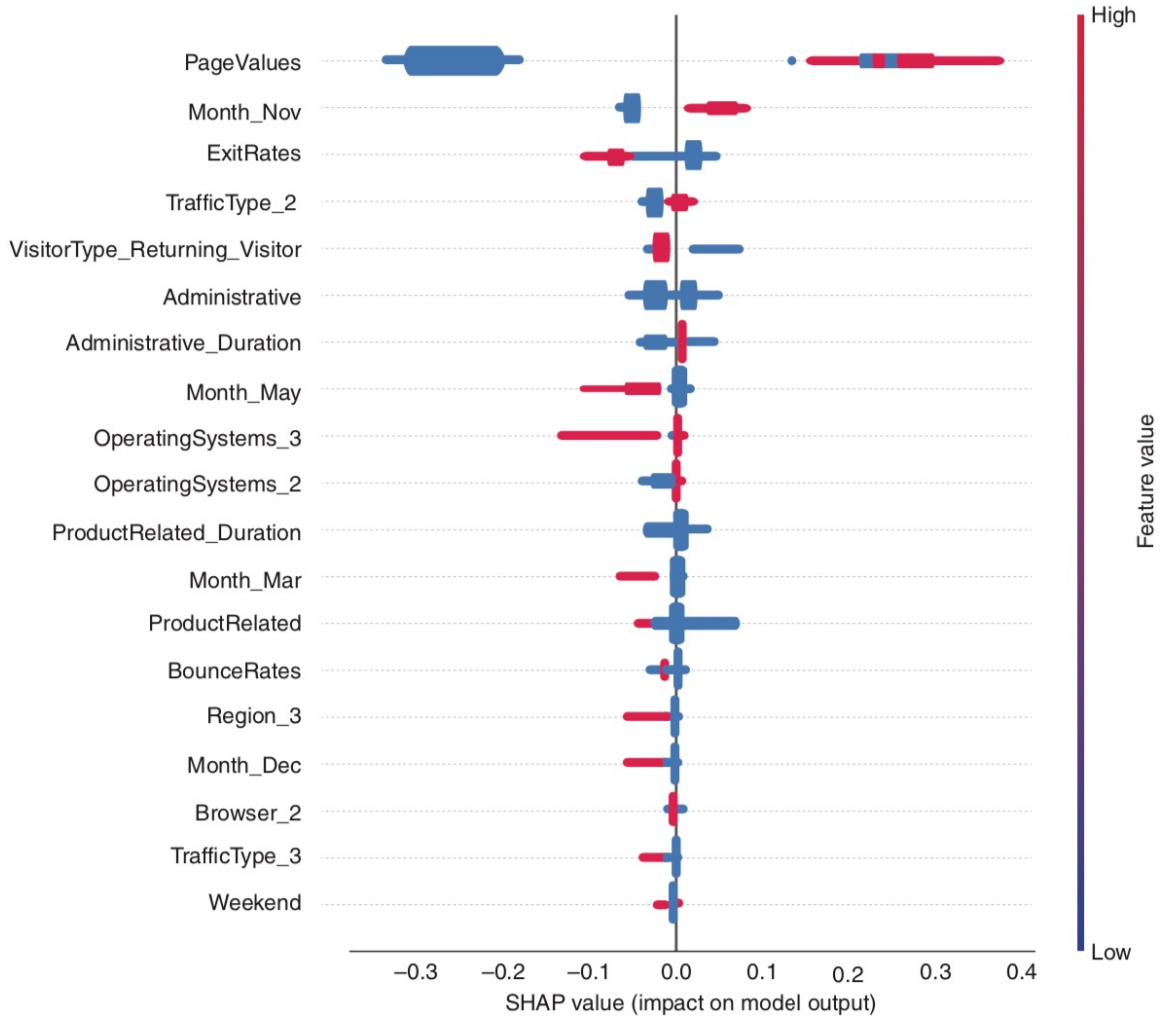


Figure 3.42 Summary plot- SHAP Global Interpretation.

- Figure 3.42 is a Summary plot, horizontal axis (*x*-axis) is the SHAP values from –0.3 to 0.4.
- All points which are below 0 will negatively impact the prediction.
- All points greater than 0 will aid prediction positively; that is, it makes the likelihood of an observation towards “buy”.

Global Interpretation

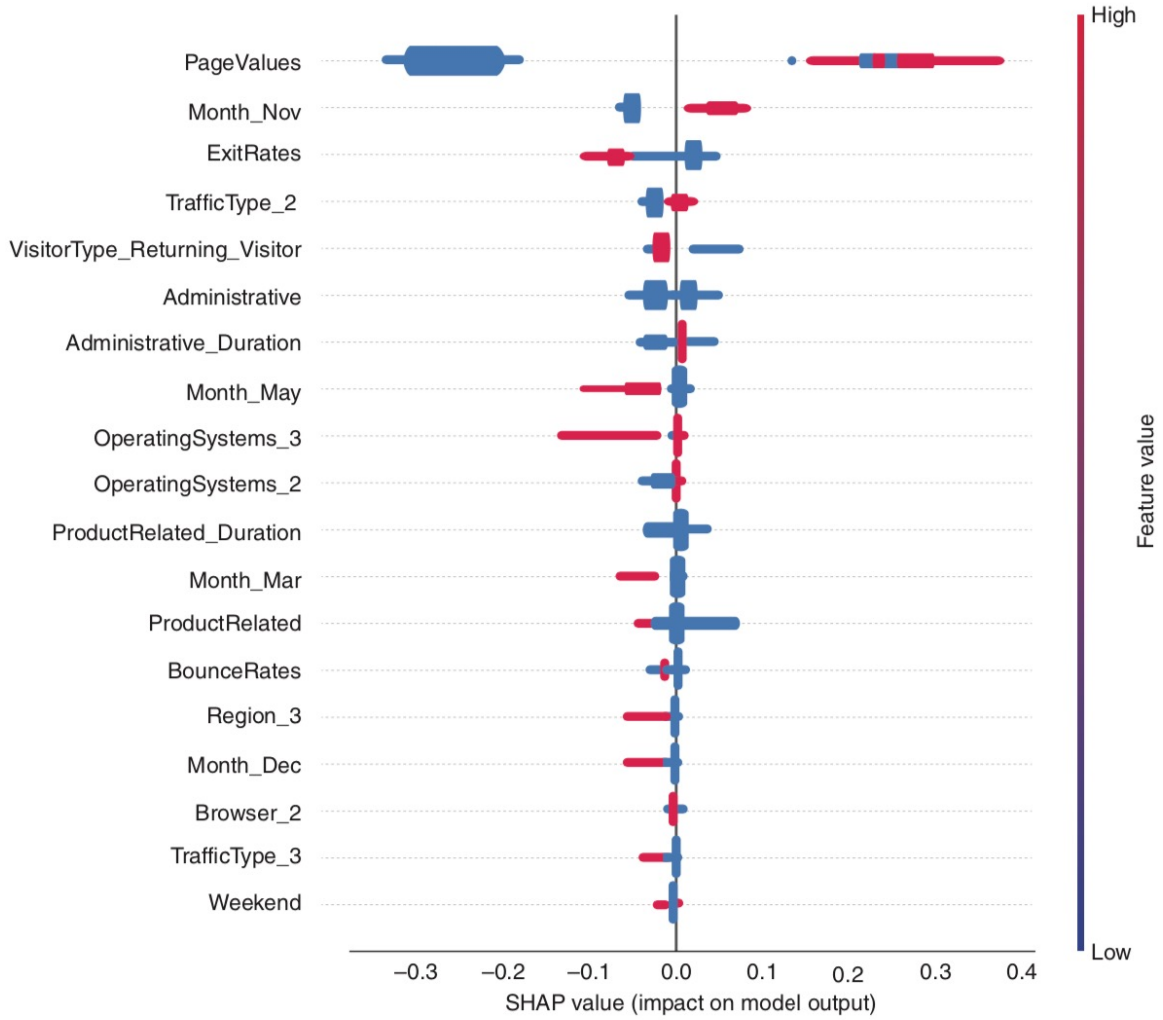


Figure 3.42 Summary plot- SHAP Global Interpretation.

- On the vertical axis (*y*-axis) are the features used in the model with colour axis depicted on the right side—blue corresponds to a small value and red to a high value.
- When the page values feature has low values (blue), it clearly decreases the likelihood of an observation being “buy”.
- As the value of the page value is high (red), the likelihood of an observation being “buy” is more.

Global Interpretation

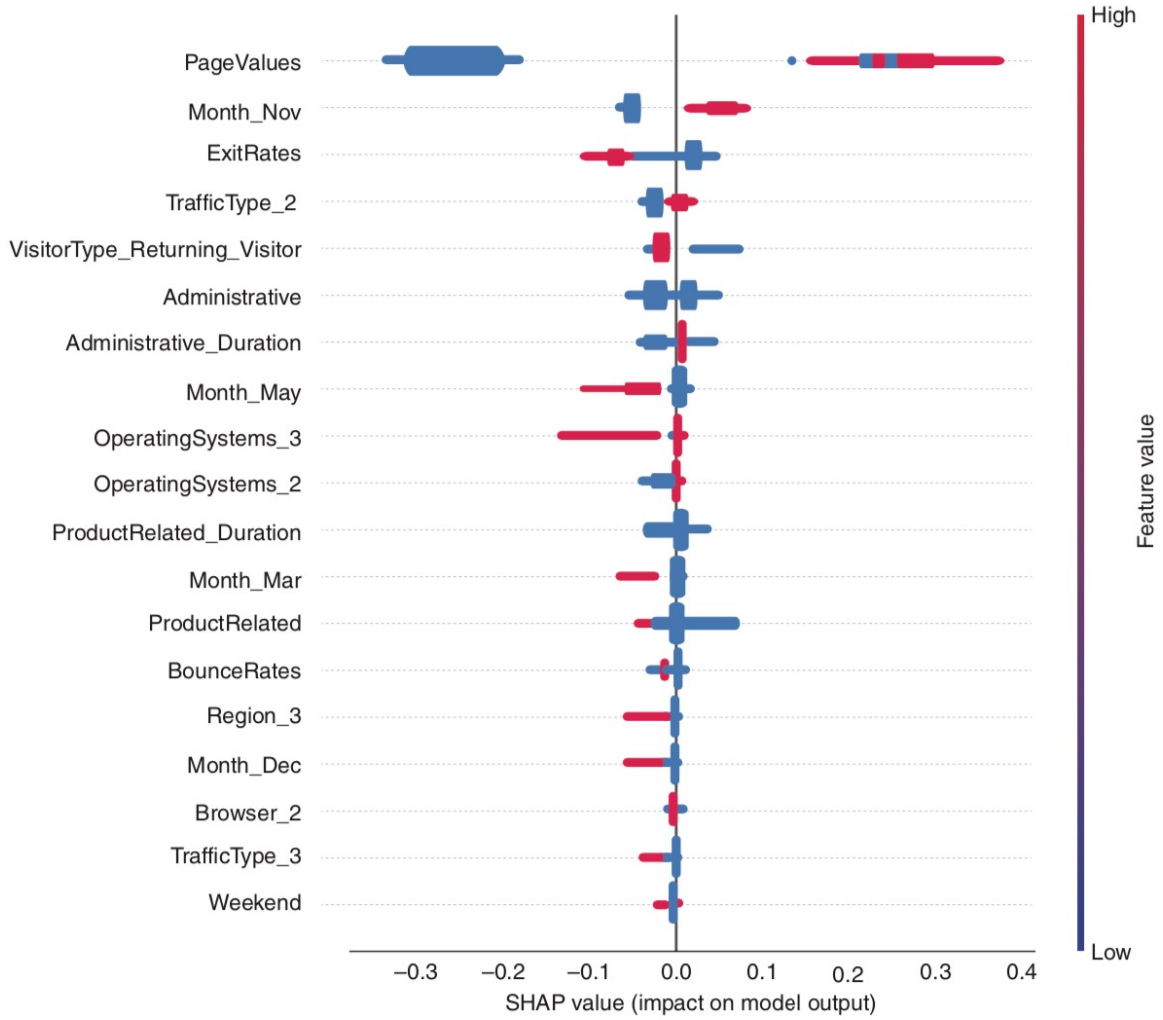


Figure 3.42 Summary plot- SHAP Global Interpretation.

- We can also see that there is a clear demarcation for feature "Month_Nov" When the values of this feature are low (blue colour), the likelihood of "not buy" is more and when the value of this feature is high (red colour), the likelihood of "buy" is more.
- So, in one single chart, we not only see how important a feature is, but also in what direction it pushes the record.

Global Interpretation

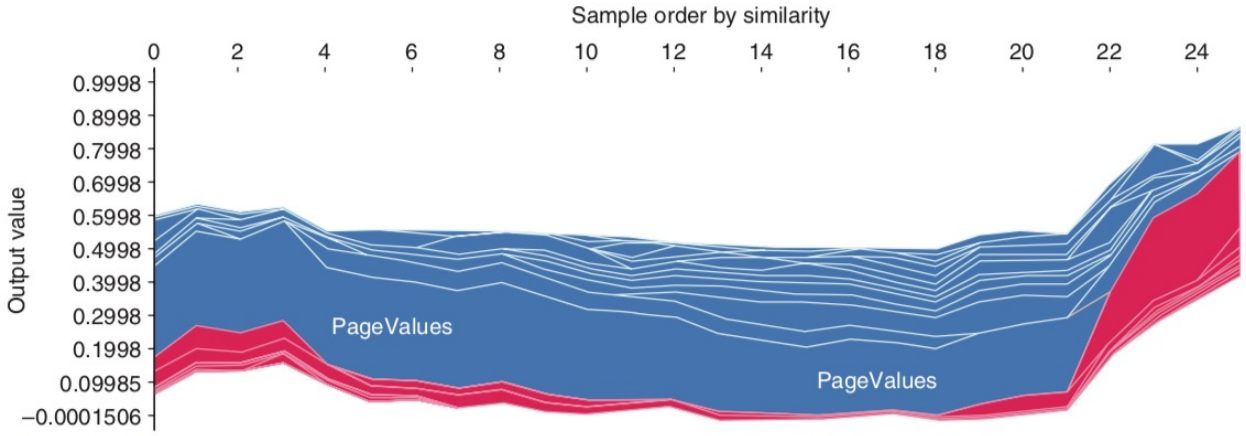


Figure 3.43 Force plot- SHAP Global Interpretation.

- Force plot is shown in Fig. 3.43.
- It is an interactive plot which can help us understand how a feature changes the likelihood of the probability of the outcome.
- The records are grouped and ordered by similarity.
- We can see that the probability of outcome being “buy” increases as page values increase.

Global Interpretation

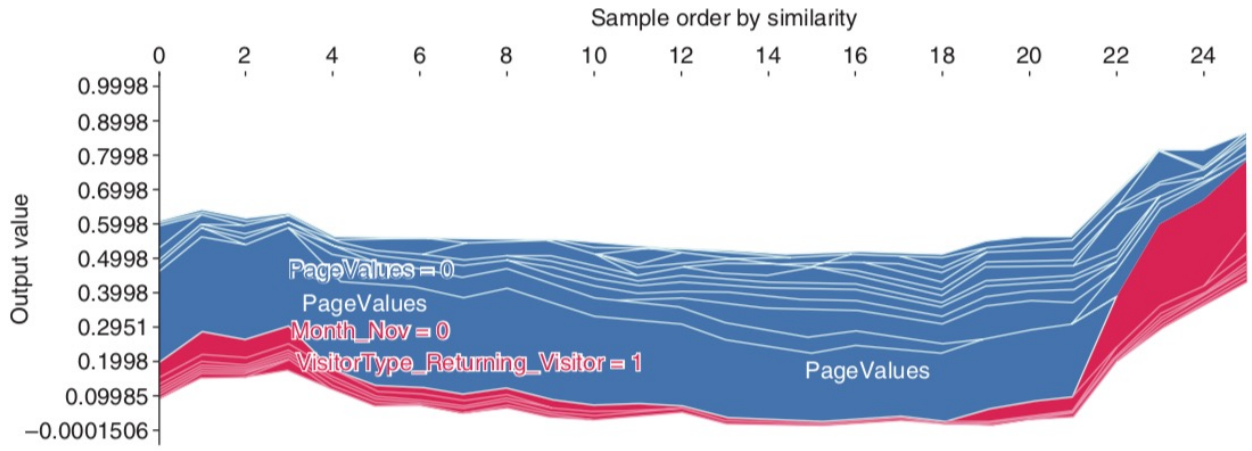


Figure 3.44 Force plot- SHAP Global Interpretation.

- It is an interactive plot and as we hover on the plot, we can see the values of all features for similar records grouped together (Fig. 3.44).

Global Interpretation

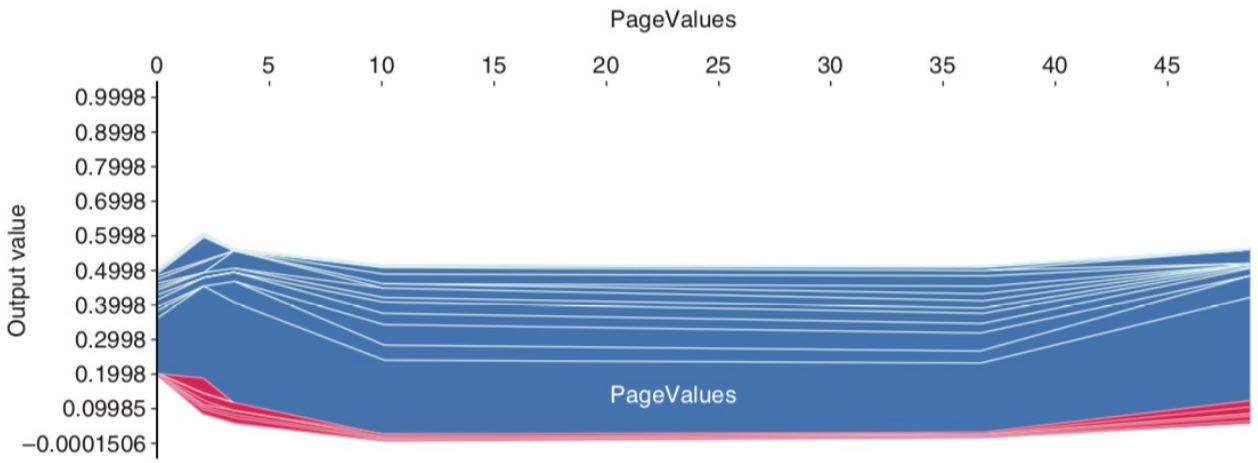


Figure 3.45 Force plot- SHAP Global Interpretation.

- We can also interact with the plot by selecting a particular feature from the horizontal (*x*-axis) drop down and see how the value of this feature affects the likelihood of the outcome.
- In Fig. 3.45, *x*-axis is all the values of page values feature from 0–50.
- We can see that the red areas, which increase the likelihood of a customer intent to “buy”, are grouped in the values near 2 to 5, and then later around 35 to 50.

Business Operation Dashboard

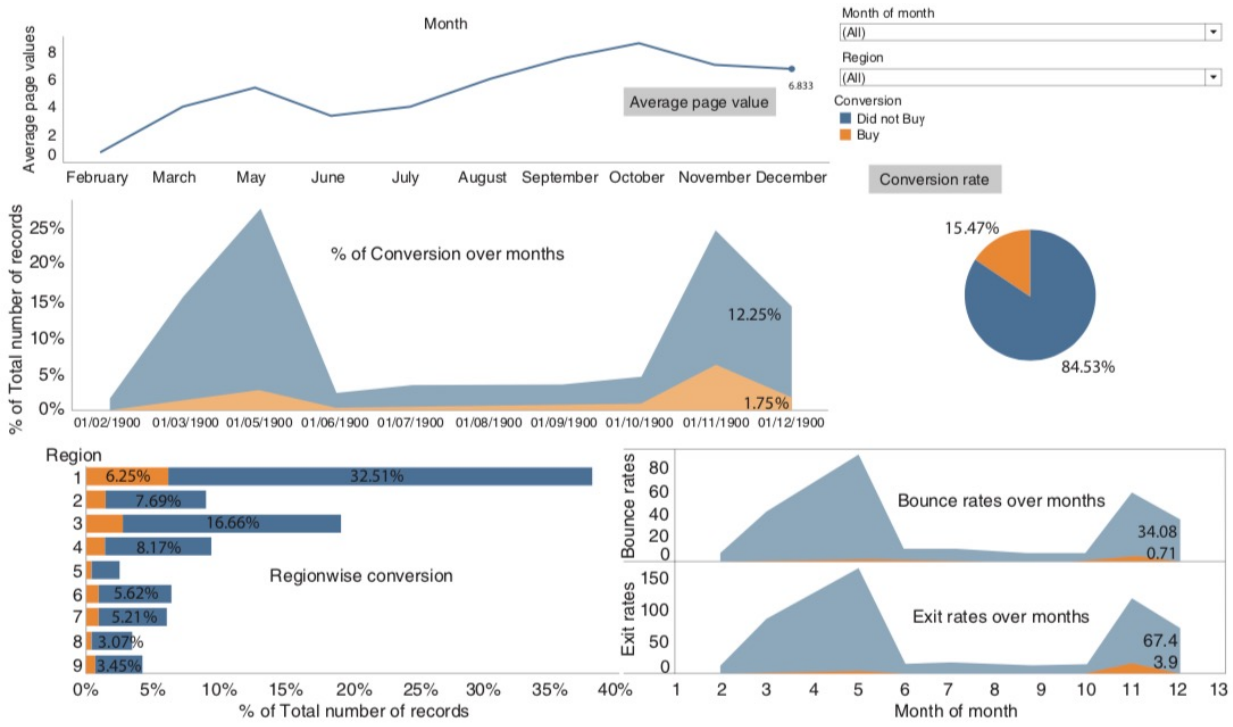


Figure 3.46 Interactive business operations dashboard.

- Interpreting the model, business/operations teams can focus on improving mobility between pages.
- Page Value is one of the important features helping the team determine when browsing customers will make a purchase.
- Also, there are certain months like November when the frequency of the purchase is high.
- This means we can capitalize on these months with different marketing strategies to aid conversion.
- We can also bucket the outcome probabilities into multiple buckets, which in turn can be used to group our customers and provide personalized marketing strategies.

References

- [Nielsen, 2005] – J. Nielsen (2005), “The Slow Tail: *Time Lag Between Visiting and Buying*”, Nielsen Norman Group, September 2005.
- [Kumar, 2021] – U. D. Kumar (2021), “Business Analytics: *The Science of Data – Driven Decision Making* ”, January 2021, ISBN: 9788126568772.
- “Comparisons, *The Data Visualization Catalogue*”, available at https:// datavizcatalogue.com/search/comparisons.html, last accessed June 4, 2020.
- “The Python Graph Gallery”, available at https://python-graph- gallery.com/, last accessed June 4, 2020.
- Dataman (2019), “*Explain model using SHAP*”, September 2019, available at https://towardsdatascience.com/ex- plain-your-model-with-the-shap-values-bc36aac4de3d, last accessed June 4, 2020.

Thank You!