

Higher Nationals - Summative Assignment Feedback Form

Student Name/ID	Ngo Tri Tai/GCD210330					
Unit Title	1641 – Business Intelligence					
Assignment Number	2	Assessor	Nguyen The Nghia			
Submission Date	25/8/2024	Date Received 1st submission	25/8/2024			
Re-submission Date		Date Received 2nd submission				
Grading grid						
P1	P2	M1	M2	D1	D2	
Assessor Feedback:						
<p>*Please note that constructive and useful feedback should allow students to understand:</p> <ul style="list-style-type: none"> a) Strengths of performance b) Limitations of performance c) Any improvements needed in future assessments <p>Feedback should be against the learning outcomes and assessment criteria to help students understand how these inform the process of judging the overall grade.</p> <p>Feedback should give full guidance to the students on how they have met the learning outcomes and assessment criteria.</p>						
Grade:	Assessor Signature:			Date:		
Resubmission Feedback:						
*Please note resubmission feedback is focussed only on the resubmitted work						
Grade:	Assessor Signature:			Date:		

Internal Verifier's Comments:
Signature & Date:

* Please note that grade decisions are provisional. They are only confirmed once internal and external moderation has taken place and grades decisions have been agreed at the assessment.

STUDENT ASSESSMENT SUBMISSION AND DECLARATION

When submitting evidence for assessment, each student must sign a declaration confirming that the work is their own.

Student name: Ngo Tri Tai		Assessor name: Nguyen The Nghia	
Issue date: 25/8/2024	Submission date: 25/8/2024	Submitted on: CMS	
Programme: BTEC			
Unit: 1641 – Business Intelligence			
Assignment number and title: 2			

Plagiarism

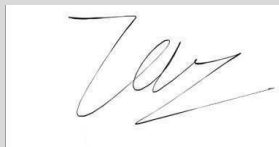
Plagiarism is a particular form of cheating. Plagiarism must be avoided at all costs and students who break the rules, however innocently, may be penalised. It is your responsibility to ensure that you understand correct referencing practices. As a university level student, you are expected to use appropriate references throughout and keep carefully detailed notes of all your sources of materials for material you have used in your work, including any material downloaded from the Internet. Please consult the relevant unit lecturer or your course tutor if you need any further advice.

Student Declaration

Student declaration

I certify that the assignment submission is entirely my own work and I fully understand the consequences of plagiarism. I declare that the work submitted for assessment has been carried out without assistance other than that which is acceptable according to the rules of the specification. I certify I have clearly referenced any sources and any artificial intelligence (AI) tools used in the work. I understand that making a false declaration is a form of malpractice.

Student signature:



Date: 11/8/2024

TABLE OF CONTENTS

A. Understanding Business Intelligence and its Tools (P3)	7
I. Introduction to Business Intelligence (BI)	7
1. Definition of BI	7
2. How does the Business Intelligence process work	7
3. Why business intelligence is important	8
4. Examples of BI Application in Businesses	8
II. Business Intelligence Techniques	12
1. Data Mining	12
2. Data Warehousing	13
3. Data Visualization	13
4. Predictive Analytics	13
5. Advanced Statistical Analysis	13
III. Business Intelligence Tools	14
1. Datapine	14
2. Domo	15
3. Microsoft Power BI	16
4. Tableau	16
5. Zoho Analytics	17
B. Designing a Business Intelligence Tool for Decision Support (P4)	18
I. Description of the dataset used for the project	18
II. Pre-processing of Data	19
1. General Data Description	19
2. Statistical Data Description	20

3. Missing Data Handling	23
4. Data Consistency Check	24

TABLE OF FIGURES

Figure 1: American Express	9
Figure 2: Coca-Cola	10
Figure 3: StarBucks	11
Figure 4: Netflix	12
Figure 5: Datapine	14
Figure 6: Domo	15
Figure 7: Microsoft Power BI	16
Figure 8: Tableau	16
Figure 9: Zoho Analytics	17
Figure 10: Car Sales Dataset	18
Figure 11: Displaying the First 10 Rows.....	20
Figure 12: Displaying the Last 10 Rows.....	21
Figure 13: Overview of the dataset's structur.....	22
Figure 14: Basic statistics for numerical column.....	23
Figure 15: Statistical information for columns.....	24
Figure 16: Calculating Percentage of Missing Values	25
Figure 17: Check for Incorrect Values or Formats	25

A. Understanding Business Intelligence and its Tools (P3):

I. Introduction to Business Intelligence (BI):

1. Definition of BI:

- Executives, managers, and staff can make educated business decisions with the help of business intelligence (BI), a technology-driven method that analyzes data and delivers insightful information. Organizations obtain data for analysis, prepare it for analysis, run queries on the data, and produce data visualizations, BI dashboards, and reports as part of the business intelligence (BI) process. The data can come from both internal and external IT systems. Business users can obtain analytics results through these outputs for the purposes of strategic planning and operational decision-making.
- Enhancing business decisions is the main goal of BI efforts; doing so will boost revenue, increase operational effectiveness, and provide a competitive advantage over competitors. In order to achieve this goal, business intelligence (BI) combines a wide range of analytical, data management, and reporting technologies with different approaches to data management and analysis.

2. How does the business intelligence process work:

- There are numerous crucial steps in the business intelligence process:
 - + **Data Storage:** Smaller data marts or data warehouses are used to store business intelligence data. Data marts hold information that is tailored to certain departments or business units, whereas data warehouses are made to store data for the entire enterprise. Diverse data kinds, including unstructured and semistructured data, are also stored in data lakes, which are built on big data systems like Hadoop.
 - + **Data Integration and Cleansing:** Raw data from several source systems is combined, cleaned, and integrated before being used in BI applications. This guarantees the data's consistency and accuracy. Tools for data integration and quality control are used for this.
 - + **Data Preparation:** To prepare them for analysis, data sets are modeled and arranged in this step. To make the data acceptable for querying and analysis, it must be transformed and structured.
 - + **Analytical Querying:** Following preparation of the data, analytical queries are run on it. This entails executing queries to draw particular conclusions and data-specific information.
 - + **Distribution of Findings:** Business users receive other findings as well as key performance indicators (KPIs). This can be accomplished by giving users access to and interpretation of the data through reports, dashboards, or data visualizations.

+ **Decision-Making:** Business choices are influenced and guided by the data that business intelligence provides. The knowledge gleaned from the data analysis is put to use by managers, executives, and other stakeholders in order to plan and make decisions.

3. Why Business Intelligence is important:

- Business intelligence is essential for optimizing an organization's business processes through the utilization of pertinent data. Through efficient application of business intelligence (BI) technologies and methodologies, organizations can get significant insights from their data, facilitating more informed choices concerning business procedures and tactics. This data-driven approach to decision-making results in larger profits in the end since it boosts revenue, accelerates business growth, and improves productivity.
- Organizations largely depend on elements like acquired knowledge, prior experiences, intuition, and gut instincts when making crucial business decisions in the absence of business intelligence integration. These approaches might produce good results, but since they lack data-driven backing, they are prone to mistakes and blunders. By giving decision-makers access to a solid data base, business intelligence closes this gap and enables them to make more confident and accurate decisions.

4. Examples of BI Application in Businesses:

- Business intelligence (BI) is used by the wealthiest companies to boost sales, cultivate repeat business, simplify processes, improve advertising campaigns, raise capital, predict consumer behavior, and find new business opportunities.

* **American Express:**



Figure 1: American Express

- As American Express has shown, business intelligence is essential to the financial industry. The business uses this technology well to provide new payment options and customize offers for clients. American Express has been able to determine that up to 24 percent of Australian consumers are likely to close their accounts within four months as a result of its market trials in Australia. Equipped with this understanding, the business adopts preemptive actions to keep these clients. Furthermore, by using BI, American Express can more effectively identify fraudulent activity and protect consumers whose card information may have been compromised, strengthening their overall security protocols.

*** Coca-Cola:**



Figure 2: Coca-Cola

- With 35 million Twitter followers and an astounding 105 million Facebook likes, Coca-Cola leverages its massive social media presence to get important social media data. The business is able to detect instances in which images of their beverages are uploaded on the internet by utilizing AI-powered image recognition technology. Coca-Cola gains vital insights into the demographics and geographic regions of its customers as well as the motivations behind online brand mentions by combining this data with the powers of business intelligence. With this data at hand, the business may present more precisely tailored adverts to the audience, which boosts click-through rates four times higher than those of generic commercials.

* **Starbucks:**



Figure 3: StarBucks

- Starbucks leverages the strength of its extensively used mobile application and loyalty card program, which provide the business access to a vast amount of individual purchase data from millions of customers. Starbucks uses business intelligence technologies to leverage this important information in order to predict client purchases and send out personalized offers via email and app that are catered to the individual's tastes. This focused strategy successfully draws in current customers by encouraging them to visit Starbucks locations more regularly, which boosts sales and cultivates a devoted following.

* **Netflix:**



Figure 4: Netflix

- With 148 million subscribers, Netflix has a big advantage in business analytics because to its large user base. By examining user behavior and preferences, the organization formulates and validates novel programming concepts, one of the many ways it efficiently uses data. Netflix uses business analytics as well in order to boost user interaction with their content. To the degree that it powers over 80% of the streamed material, the platform's recommendation system is incredibly skilled at promoting targeted content.

II. Business Intelligence Techniques:

- Techniques for gathering, analyzing, and interpreting data to support decision-making and enhance company outcomes are referred to as business intelligence techniques.

1. Data Mining:

- A crucial step in the business intelligence process, data mining is the extraction of useful knowledge from raw data. It is particularly useful for transactional data analysis. Businesses can use this technique to obtain detailed insights into a variety of topics, such as staff performance, product sales, and customer demographics. Data mining is an incredibly useful business intelligence technique, especially in sectors like retail, banking, healthcare, and hospitality that mostly rely on transactional data. Data mining helps firms find previously overlooked problems and obtain a better knowledge of their operations by

locating buying trends and patterns, discovering customer service issues, and uncovering operational faults.

2. Data Warehousing:

- The process of data warehousing entails collecting unprocessed data from many sources, including sensors and Internet of Things devices, and storing it in a primary database. A predetermined structure is usually included in the architecture of this database, which allows for multiple query options to be used to extract pertinent data.
- Data warehousing is frequently used by data-intensive businesses, such as social media networks and internet service providers, to manage massive amounts of data. Organizations may gather and store vast volumes of data thanks to the expansive and scalable nature of these data warehouses. The basic goal of data warehousing is to keep the data in an easily accessible manner so that additional research and insights, including consumer demographics and spending habits, may be extracted.

3. Data Visualization:

- Data visualization is the process of transforming data into aesthetically pleasing graphs and charts to improve its understandability for non-technical people. This business intelligence method seeks to improve non-technical users' access to raw data.
- Data visualization is a useful tool for learning about a variety of topics, such as employee performance and product sales. Difficult and convoluted data sets can be simplified into information that is easily understood by using data visualization techniques. This method helps management and executives who might be feeling overpowered by the volume of data accessible to them because it makes interpretation easier and allows for well-informed decision-making.

4. Predictive Analytics:

- A sophisticated statistical analysis technique called predictive analytics uses past data to forecast future events. Marketing uses it extensively to find possible clients by looking at their demographics and behavior. Moreover, predictive analytics finds use in sectors like healthcare and insurance. Organizations looking to maximize their marketing efforts through customized offerings and targeted campaigns will find it very beneficial. Predictive analytics also assists with risk identification and mitigation, lowering the probability that a risk will materialize. In general, this method helps businesses to make data-driven decisions that improve company outcomes and to reach their full marketing potential.

5. Advanced Statistical Analysis:

- Using complex mathematical formulas, advanced statistical analysis is a business intelligence method that finds patterns in data. It works best when paired with additional business intelligence strategies.

Organizations may learn more from their data and uncover patterns and connections that they may have missed otherwise by applying sophisticated statistical analysis.

III. Business Intelligence Tools:

- Software programs or platforms known as business intelligence tools are made to collect, process, and present data in order to aid in organizational decision-making. From a variety of data sources, these tools offer users a range of features to extract insights, generate reports, and build interactive visualizations.

1. Datapine:

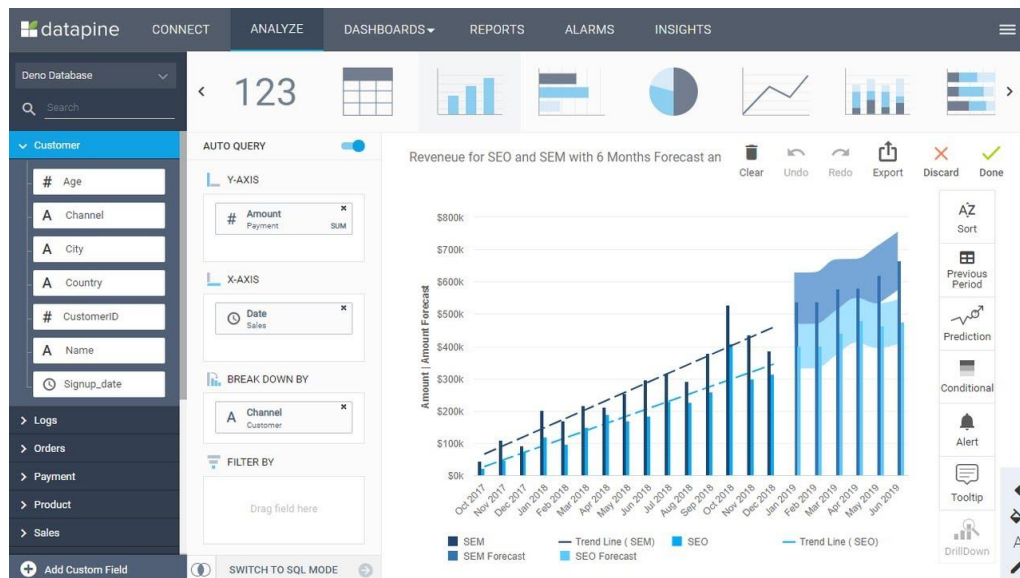


Figure 5: Datapine

- A complete business intelligence platform called Datapine was created to make data analytics easier to understand for non-technical people. Datapine's self-service analytics methodology facilitates the seamless integration of various data sources, advanced data analysis, interactive dashboard creation, and the extraction of actionable insights for business objectives for both data analysts and business users.

2. Domo:

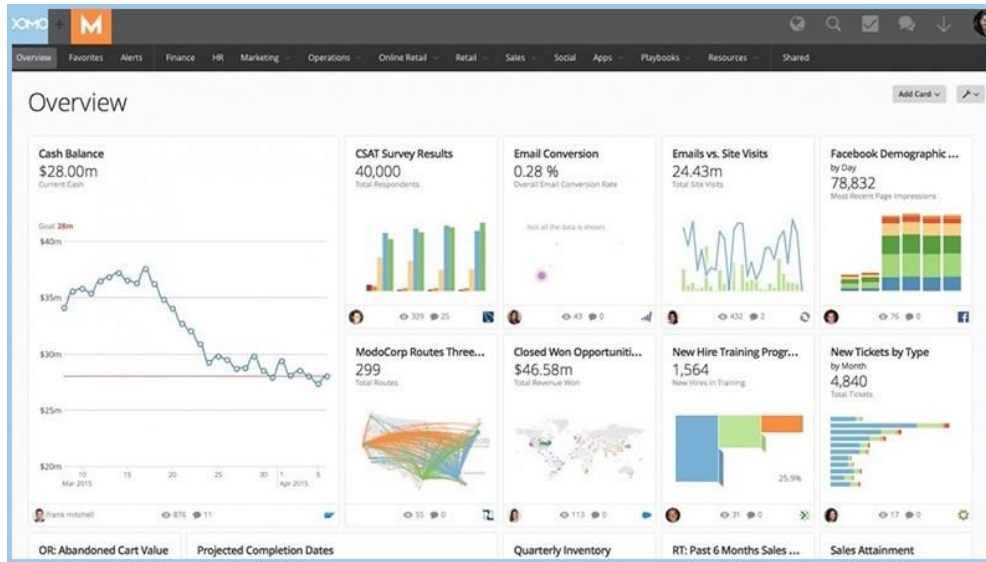


Figure 6: Domo

- Spreadsheets, databases, and social media are just a few of the data sources that Domo's entirely cloud-based business intelligence platform effortlessly connects. Both small and huge international enterprises can have their demands met by it. The platform offers broad visibility and macro and micro-level analysis capabilities. It provides predictive analysis capabilities by utilizing Mr. Roboto, its AI engine. Users may calculate marketing ROI across several channels and obtain critical details like cash balances and regional sales performance with Domo. There might be a learning curve to using the platform, though, and users may run into difficulties when downloading analyses from the cloud for personal use.

3. Microsoft Power BI:

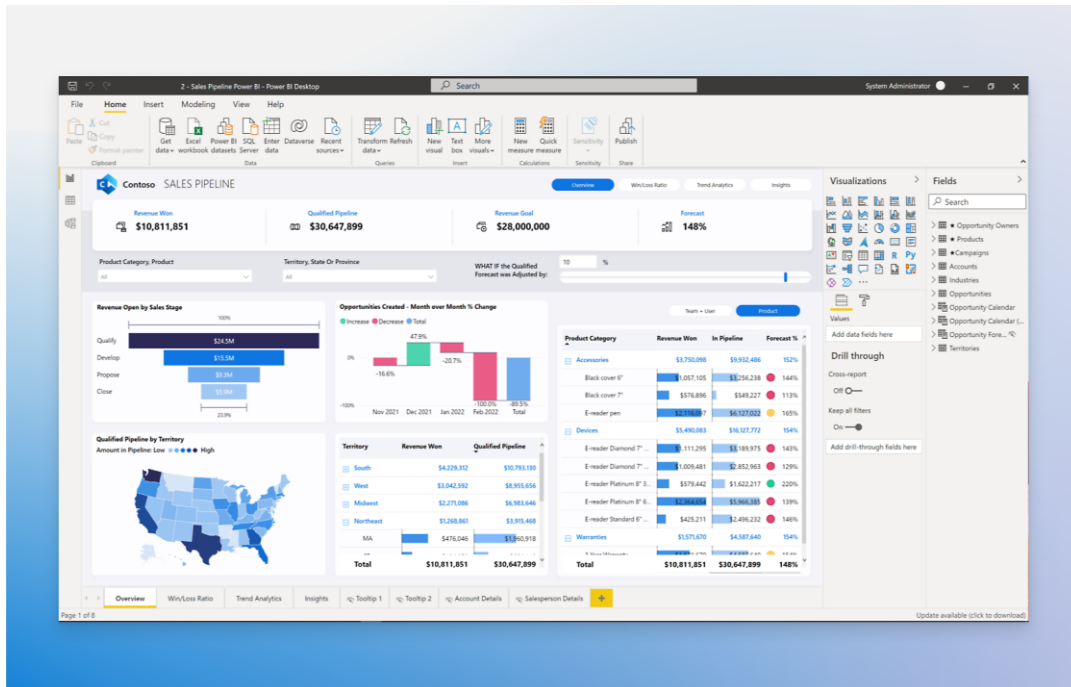


Figure 7: Microsoft Power BI

- The web-based business analytics toolkit Microsoft Power BI is well-known for its outstanding data visualization features. It gives users the ability to spot patterns in real time and offers fresh connectors that improve the effectiveness of campaigns. Because Microsoft Power BI is web-based, it can be accessed from almost anywhere. Users may deliver reports and real-time dashboards with ease, as well as integrate their apps smoothly.

4. Tableau:

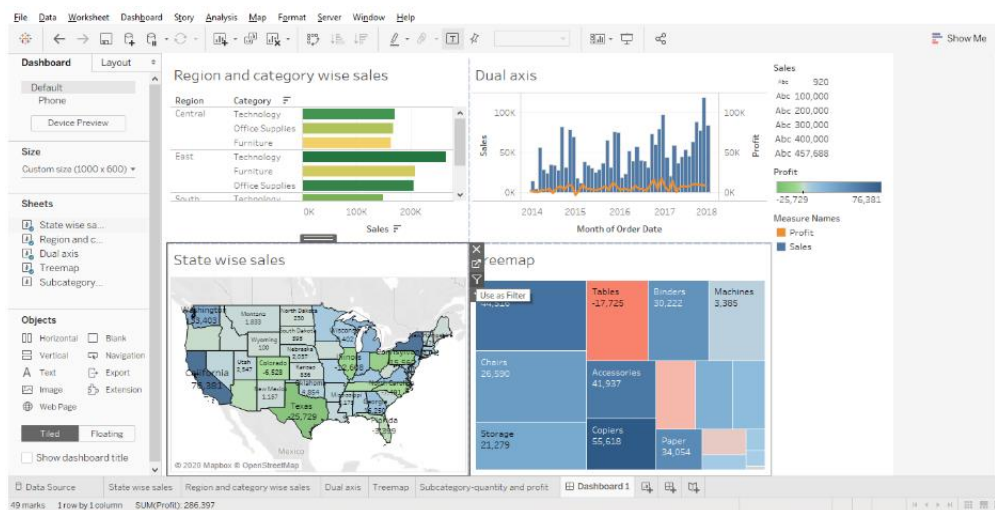


Figure 8: Tableau

- Data visualization and discovery are the areas of expertise for the business intelligence application Tableau. Without requiring IT assistance, it lets users analyze, visualize, and share data with ease. Many data sources, such as Google Analytics, Salesforce, Oracle, MS SQL, and Microsoft Excel, are supported by Tableau. Dashboards that are easy to use and have a good design can help users explore data more effectively. Tableau also provides a range of stand-alone products, including Tableau Desktop (ideal for individuals), Tableau Server (an analytics tool for businesses that can be run locally), and Tableau Online (an analytics tool that can be hosted for businesses), among others.

5. Zoho Analytics:



Figure 9: Zoho Analytics

- An excellent business intelligence tool, Zoho Analytics excels at offering thorough reporting and in-depth data analysis capabilities. It allows you to arrange regular periods for data updates and supports automatic data synchronization. With integration APIs, users may quickly and easily build connectors to combine and aggregate data from several sources. The application features an easy-to-use editor that lets users create customized dashboards and reports, freeing them up to concentrate on the important details. To help team members collaborate easily, Zoho Analytics additionally has a separate commenting box in its sharing choices.

B. Designing a Business Intelligence Tool for Decision Support (P4):

I. Description of the dataset used for the project:

- The Car Sales Dataset is an extensive set of transactional data that provides insightful information on the trends and purchasing habits of consumers. Date, Customer Name, Gender, Annual Income, Dealer Name, Company, Model, Engine, Transmission, Color, Price(\$), Body Style, and Dealer Region are among the crucial information it includes. Businesses can improve their marketing tactics and maximize their product offers by analyzing this dataset to obtain a better understanding of the buying habits of their customers, identify market trends, and make well-informed decisions. This dataset is very useful for studying and comprehending the dynamics of auto sales.

Car_id	Date	Customer	Gender	Annual Inc	Dealer_Na	Company	Model	Engine	Transmissi	Color	Price (\$)	Dealer_Nc	Body Style	Phone	Dealer_Region
C_CND_0C	1/2/2022	Geraldine	Male	13500	Buddy Stoi	Ford	Expedition	Double A A	Auto	Black	26000	06457-383	SUV	8264678	Middletown
C_CND_0C	1/2/2022	Gia	Male	1480000	C & M Mo	Dodge	Durango	Double A A	Auto	Black	19000	60504-711	SUV	6848189	Aurora
C_CND_0C	1/2/2022	Gianna	Male	1035000	Capitol Kl	Cadillac	Eldorado	Overhead	Manual	Red	31500	38701-804	Passenger	7298798	Greenville
C_CND_0C	1/2/2022	Giselle	Male	13500	Chrysler of	Toyota	Celica	Overhead	Manual	Pale White	14000	99301-388	SUV	6257557	Pasco
C_CND_0C	1/2/2022	Grace	Male	1465000	Chrysler Pl	Acura	TL	Double A A	Auto	Red	24500	53546-942	Hatchback	7081483	Janesville
C_CND_0C	1/2/2022	Guadalupe	Male	850000	Classic Chi	Mitsubishi	Diamante	Overhead	Manual	Pale White	12000	85257-31C	Hatchback	7315216	Scottsdale
C_CND_0C	1/2/2022	Hailey	Male	1600000	Clay Johns	Toyota	Corolla	Overhead	Manual	Pale White	14000	78758-784	Passenger	7727879	Austin
C_CND_0C	1/2/2022	Graham	Male	13500	U-Haul CO	Mitsubishi	Galant	Double A A	Auto	Pale White	42000	78758-784	Passenger	6206512	Austin
C_CND_0C	1/2/2022	Naomi	Male	815000	Rabun Use	Chevrolet	Malibu	Overhead	Manual	Pale White	82000	85257-31C	Hardtop	7194857	Pasco
C_CND_0C	1/2/2022	Grayson	Female	13500	Rabun Use	Ford	Escort	Double A A	Auto	Pale White	15000	85257-31C	Passenger	7836892	Scottsdale
C_CND_0C	1/2/2022	Gregory	Male	13500	Race Car F	Acura	RL	Overhead	Manual	Pale White	31000	78758-784	SUV	7995489	Austin
C_CND_0C	1/2/2022	Amar'E	Male	13500	Race Car F	Nissan	Pathfinder	Double A A	Auto	Pale White	46000	78758-784	Hardtop	7288103	Pasco
C_CND_0C	1/2/2022	Griffin	Male	885000	Saab-Belle	Mercury	Grand Mar	Double A A	Auto	Black	9000	60504-711	SUV	6842408	Aurora
C_CND_0C	1/2/2022	Harrison	Male	13500	Scrivener F	BMW	323i	Double A A	Auto	Pale White	15000	38701-804	Hatchback	7558767	Greenville
C_CND_0C	1/2/2022	Zainab	Male	722000	Buddy Stoi	Chrysler	Sebring Co	Overhead	Manual	Pale White	26000	06457-383	Sedan	7677191	Middletown
C_CND_0C	1/2/2022	Zara	Male	746000	C & M Mo	Subaru	Forester	Overhead	Manual	Pale White	17000	60504-711	Hatchback	8431908	Aurora
C_CND_0C	1/2/2022	Zoe	Female	535000	Capitol Kl	Hyundai	Accent	Overhead	Manual	Black	18000	38701-804	Hatchback	7814646	Greenville
C_CND_0C	1/2/2022	Zoe	Female	570000	Chrysler of	Cadillac	Eldorado	Double A A	Auto	Pale White	31000	99301-388	Passenger	7456650	Pasco
C_CND_0C	1/2/2022	Aaliyah	Male	685000	Chrysler Pl	Toyota	Land Cruis	Double A A	Auto	Pale White	33000	53546-942	SUV	7627010	Janesville
C_CND_0C	1/2/2022	Abigail	Male	455000	Classic Chi	Honda	Accord	Double A A	Auto	Pale White	21000	85257-31C	Sedan	6736704	Scottsdale
C_CND_0C	1/2/2022	Adrianna	Male	13500	Clay Johns	Toyota	4Runner	Overhead	Manual	Black	25000	78758-784	Sedan	7889827	Austin
C_CND_0C	1/2/2022	Joshua	Male	2500000	Classic Chi	Infiniti	I30	Double A A	Auto	Black	21000	85257-31C	Hardtop	6183219	Austin
C_CND_0C	1/2/2022	Marcus	Male	585000	Diehl Moti	Audi	A4	Overhead	Manual	Pale White	12000	06457-383	Hardtop	8097778	Middletown
C_CND_0C	1/2/2022	Arthur	Male	920000	Star Enterj	Porsche	Carrera Ca	Double A A	Auto	Pale White	18000	99301-388	Passenger	7959858	Pasco
C_CND_0C	1/2/2022	Lizzie	Male	672000	Suburban f	Volkswage	Jetta	Double A A	Auto	Pale White	22000	53546-942	Passenger	8570849	Janesville
C_CND_0C	1/2/2022	Florian	Male	801250	Tri-State N	Dodge	Viper	Double A A	Auto	Pale White	31250	85257-31C	SUV	8520534	Scottsdale
C_CND_0C	1/2/2022	Cassandra	Female	820000	U-Haul CO	Buick	Regal	Double A A	Auto	Black	19000	78758-784	Passenger	6362556	Austin
C_CND_0C	1/2/2022	Srielle	Male	791000	Progressiv	Chrysler	LHS	Overhead	Manual	Pale White	41000	53546-942	Hatchback	6281210	Janesville

Figure 10: Car Sales Dataset

- + **Car_id**: a distinct number for every vehicle in the collection.
- + **Date**: Date of the vehicle sale agreement.
- + **Customer Name**: Name of the buyer of the vehicle.
- + **Gender**: Customer's gender (e.g., Male, Female).
- + **Annual Income**: The client's annual income.
- + **Dealer_Name**: Name of the participating auto dealer in the transaction.

- + **Company:** automobile manufacturer or brand.
- + **Model:** The vehicle's model name.
- + **Engine:** details on the engine of the vehicle.
- + **Transmission:** the vehicle's gearbox type (e.g., Automatic, Manual).
- + **Color:** exterior color of the automobile.
- + **Price(\$):** The automobile for sale's listed price.
- + **Body Style:** The body style or design of the car (SUV, sedan, etc.).
- + **Dealer_Region:** Geographical area or the auto dealer's location.

II. Pre-processing of Data:

1. General Data Description:

- It's important to have a general idea of the dataset before diving into the details of the data processing techniques. This entails understanding the data's structure, the importance of each column, and how it is organized. This first phase helps us become acquainted with the dataset we are dealing with and ensures that we are sufficiently ready for the next steps of data processing.
- We use the Pandas package to efficiently preprocess datasets in our Python environment within Visual Studio Code, which streamlines data analysis and manipulation.

* **Loading the Dataset:**

```

43 import pandas as pd
44 import plotly
45 import plotly.express as px
46 import plotly.graph_objects as go
47 import statsmodels.api as sm
48 from plotly.subplots import make_subplots
49 df = pd.read_csv("/Program Files/Spyder/pkgs/pandas/io/Car Sales.xlsx - car_data.csv")

```

- A file named "Car Sales.xlsx - car_data.csv" is read by the code. On your computer, the file is located in the designated folder "C:\Program Files\Spyder\pkgs\pandas\io\Car Sales.xlsx - car_data.csv." The Pandas library is referred to by the "pd" portion of the code. The CSV file's contents are read using the Pandas library's "read_csv" function.

- * **Displaying the First 10 Rows:** The head() method is used to display the first 10 rows of the dataset.

```
In [4]: print(df.head(10))
```

	Car_id	Date	Customer Name	...	Body Style	Phone	Dealer_Region
0	C_CND_000001	1/2/2022	Geraldine	...	SUV	8264678	Middletown
1	C_CND_000002	1/2/2022	Gia	...	SUV	6848189	Aurora
2	C_CND_000003	1/2/2022	Gianna	...	Passenger	7298798	Greenville
3	C_CND_000004	1/2/2022	Giselle	...	SUV	6257557	Pasco
4	C_CND_000005	1/2/2022	Grace	...	Hatchback	7081483	Janesville
5	C_CND_000006	1/2/2022	Guadalupe	...	Hatchback	7315216	Scottsdale
6	C_CND_000007	1/2/2022	Hailey	...	Passenger	7727879	Austin
7	C_CND_000008	1/2/2022	Graham	...	Passenger	6206512	Austin
8	C_CND_000009	1/2/2022	Naomi	...	Hardtop	7194857	Pasco
9	C_CND_000010	1/2/2022	Grayson	...	Passenger	7836892	Scottsdale

[10 rows x 16 columns]

Figure 11: Displaying the First 10 Rows

* **Displaying the Last 10 Rows:** The tail() method is used to display the last 10 rows of the dataset.

```
In [7]: print(df.tail(10))
```

	Car_id	Date	Customer Name	...	Body Style	Phone	Dealer_Region
23896	C_CND_023897	12/31/2023	Simi	...	SUV	8744249	Aurora
23897	C_CND_023898	12/31/2023	Simone	...	Sedan	6819422	Greenville
23898	C_CND_023899	12/31/2023	Skylar	...	Hatchback	6225183	Pasco
23899	C_CND_023900	12/31/2023	Yuna	...	Hatchback	8384785	Aurora
23900	C_CND_023901	12/31/2023	Nathan	...	Sedan	8170003	Greenville
23901	C_CND_023902	12/31/2023	Martin	...	Passenger	8583598	Pasco
23902	C_CND_023903	12/31/2023	Jimmy	...	Hardtop	7914229	Middletown
23903	C_CND_023904	12/31/2023	Emma	...	Sedan	7659127	Scottsdale
23904	C_CND_023905	12/31/2023	Victoire	...	Passenger	6030764	Austin
23905	C_CND_023906	12/31/2023	Donovan	...	Hardtop	7020564	Middletown

[10 rows x 16 columns]

Figure 12: Displaying the Last 10 Rows

2. Statistical Data Description:

- For each of the dataset's numerical columns, compute the basic statistical measures. The aforementioned metrics comprise computations such as the mean, median, standard deviation, and quartile division of the data. By providing useful information about the data's central tendency and distribution, these statistics make it possible to identify any odd or anomalous numbers.

- * **df.info():** The info() method provides an overview of the dataset's composition, which includes details about the data types and the count of non-null values for every column.

```
In [8]: print(df.info())
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 23906 entries, 0 to 23905
Data columns (total 16 columns):
#   Column                Non-Null Count  Dtype
---  -
0   Car_id                23906 non-null  object
1   Date                  23906 non-null  object
2   Customer Name         23905 non-null  object
3   Gender                23906 non-null  object
4   Annual Income         23906 non-null  int64
5   Dealer_Name           23906 non-null  object
6   Company               23906 non-null  object
7   Model                 23906 non-null  object
8   Engine                23906 non-null  object
9   Transmission          23906 non-null  object
10  Color                  23906 non-null  object
11  Price ($)              23906 non-null  int64
12  Dealer_No              23906 non-null  object
13  Body Style             23906 non-null  object
14  Phone                  23906 non-null  int64
15  Dealer_Region          23906 non-null  object
dtypes: int64(3), object(13)
memory usage: 2.9+ MB
None
```

Figure 13: Overview of the dataset's structure

=> The dataset comprises 16 columns and 23,906 entries in total. The columns show varying percentages of non-null values, which suggests that there are gaps in the data. Three of the columns have the data type int64, and the remaining thirteen have object. An estimated 2.6+ MB of RAM are used by the dataset.

- * **df.describe():** For numerical columns with either an int or float data type in the dataset, the describe() method produces basic statistical statistics.

```
In [10]: print(df.describe())
```

	Annual Income	Price (\$)	Phone
count	2.390600e+04	23906.000000	2.390600e+04
mean	8.308403e+05	28090.247846	7.497741e+06
std	7.200064e+05	14788.687608	8.674920e+05
min	1.008000e+04	1200.000000	6.000101e+06
25%	3.860000e+05	18001.000000	6.746495e+06
50%	7.350000e+05	23000.000000	7.496198e+06
75%	1.175750e+06	34000.000000	8.248146e+06
max	1.120000e+07	85800.000000	8.999579e+06

```
In [11]:
```

Figure 14: Basic statistics for numerical columns

=> With an average yearly income of about \$830,430, the dataset appears to be diversified, with members coming from a range of socioeconomic backgrounds. A large standard deviation of almost \$720,064 suggests notable differences in income. The lowest income of \$10,800 is probably part-time work or entry-level pay, while the highest income of \$11,200,000 can go to investors, executives, or company owners. In terms of automotive costs, the moderate standard deviation of roughly \$14,788 suggests affordability, with a mean of about \$28,090. The \$1,200 lowest car price seen is probably associated with older or more affordable models, while the \$85,800 highest price points to luxury or premium automobiles. This dataset generally illustrates the diversity of consumer choices and economic variability.

* **df.describe(include='object')**: This part of the output presents statistical information for columns in the dataset that are categorized and have a data type of object.

```
In [11]: df.describe(include='object')
```

```
Out[11]:
```

	Car_id	Date	...	Body	Style	Dealer_Region
count	23906	23906	...	23906	23906	23906
unique	23906	612	...	5	7	7
top	C_CND_000001	9/5/2023	...	SUV	Austin	Austin
freq	1	190	...	6374	4135	4135

```
[4 rows x 13 columns]
```

Figure 15: Statistical information for columns

=> Interesting insights are revealed by the descriptive statistics for the global car sales dataset. Different cars are denoted by different car IDs; C_CND_000001 is a frequently selected option. The range of dates

points to steady sales activity, and the frequent appearance of "Thomas" in the customer base shows devoted followers. In terms of gender distribution, men predominate. Dealership insights include Chevrolet being a well-known brand and Progressive Shippers Cooperative Association being a well-known organization. The popularity of the Diamante model may be shown in its frequency. Double Overhead Camshaft engines are the most popular, and the Auto Pale White SUV combination is preferred. Finally, Austin dealers are doing well.

3. Missing Data Handling:

- Choose the appropriate course of action for the columns that have missing data. Using more sophisticated imputation techniques, removing rows with missing data, or substituting the mean or median for missing values are all viable choices. The choice of an approach that preserves the data's integrity is essential.

* Calculating Percentage of Missing Values:

```
In [12]: print(df.isnull().sum()*100/len(df))
Car_id      0.000000
Date        0.000000
Customer Name 0.004183
Gender      0.000000
Annual Income 0.000000
Dealer_Name 0.000000
Company     0.000000
Model       0.000000
Engine      0.000000
Transmission 0.000000
Color       0.000000
Price ($)   0.000000
Dealer_No   0.000000
Body Style  0.000000
Phone       0.000000
Dealer_Region 0.000000
dtype: float64
```

Figure 16: Calculating Percentage of Missing Values

=> The following data are complete: Car_id, Date, Gender, Annual Income, Dealer_Name, Company, Model, Engine, Transmission, Color, Price (\$), Body Style, and Dealer_Region (0%). Customer Name is the lone variable that lacks values, and its missing data percentage is a pitiful 0.004183%.

=> It is improbable that the Customer Name by itself will have a direct influence on automobile sales or revenue when it comes to forecasting revenue trends. The main determinants of revenue patterns are things like product attributes, pricing policies, market demand, and economic situations. These revenue-generating elements are independent of the customer's name.

4. Data Consistency Check:

- To guarantee data consistency, it is crucial to maintain the dataset's dependability and accuracy. The "Car Sales" dataset is thoroughly examined in this section to identify and address any discrepancies, enhancing the overall quality of the data for use in subsequent analyses.

*** Check for Incorrect Values or Formats:**

```
In [13]: print(df['Gender'].unique())  
['Male' 'Female']
```

Figure 17: Check for Incorrect Values or Formats

=> By carrying out thorough tests for data consistency, we reinforce the dataset's fundamental structure, boosting confidence in its dependability and enabling perceptive analyses that support informed decision-making.