**TABLE OF CONTENTS:**

1. Task 1 (Description of Big Data technologies)

* Data lake
* In-Memory Database
* Streaming Analytics
* Edge Computing
* Artificial Intelligence
* Apache Spark

1. Task 2 (Manipulation of Car Data with Tableau)

* Find out which car manufacturer produces the lowest quantity of models.
* Find out the highest average fuel economy for urban and motorway driving from the given data set.
* Find out the high and low average fuel economy from all transmission types.
* Find out which car manufacturers have 4WD (4 wheel drive) and 2WD (2 wheel drive) models and whose engine power is more than 3.5.
* Critically evaluate the strengths and weaknesses of data analytics, using Tableau and other recommended tools (up to 3 tools) which are used for data analytics.

1. References

**TASK 1:**

**DATA LAKE:**

A data lake is a storage repository that maintains a large quantity of raw data in its original format until it is required for analytics applications. Although the hierarchical dimensions and tables used in conventional data warehouses are still used, the flat design of the data lake allows for the usage of data stored mostly in files or objects. In this way, consumers may better organize, store, and make use of their data.

With regards to Hadoop, a lot of people think about "data lakes." Spread processing framework Installation’s store data in the Hadoop Distributed File System (HDFS), which is then distributed across the nodes of a Hadoop cluster for processing. Hadoop has traditionally been used to build data lakes, but cloud object storage services are increasingly being used in their place. Some NoSQL databases may also be utilized as data lakes.

Unstructured and semi-structured data may be stored in a data lake as well as structured and unstructured data. Most data warehouses are based on relational databases, which do not work well in these types of situations. Relational systems can only store structured transaction data because they need a fixed data model. No prior schema definition is required for data lakes, which allow for a wide range of schemas. Thus, it is possible for them to deal with a wide range of data kinds in diverse formats.

Consequently, many firms' data architectures now include data lakes as a critical component. Companies use them largely for big data analytics and other data science applications needing massive amounts of data and sophisticated analytics methods, such as data mining and predictive modelling.

Data scientists and analysts may identify, compile, and analyses important data from a central place called a "data lake. " That's more difficult without one. It's also more difficult for businesses to take full use of their data assets in order to make better business choices and strategies.

**IN-MEMORY DATABASE:**

In contrast to databases that store data on discs or SSDs, in-memory databases are specifically designed databases that depend on memory for data storage. To reduce reaction times, in-memory data storage eliminates the need to access drives. In-memory databases run the risk of losing data in the event of a process or server failure since all data is stored and handled only in main memory. Data in in-memory databases may be stored on disc by logging each transaction or collecting snapshots.

Gaming leaderboards, session storage, and real-time analytics all benefit from in-memory databases since they demand microsecond response times or big spikes in traffic.

A computer's primary memory, rather than a hard drive, is used to store data in an in-memory database (IMDB). Querying data on a disc takes time that may be saved by accessing data that is already stored in memory. For applications that need fast reaction times and real-time data handling, in-memory databases are ideal. The telecommunications, finance, tourism, and gambling industries all benefit from using in-memory databases. This kind of database is also known as an MMDB, real-time database (RTDB), or an in-memory database system (IMDS).

All of the data in an in-memory database is stored in the computer's random-access memory (RAM). When you run a data query, the main memory is the only place that gets accessible. Faster access to the data than a disk-based system enables.

RAM's instability is a drawback. When an in-memory database collapses, all of the data is lost. In-memory databases may benefit from the development of non-volatile random-access memory (NVRAM). Flash is a good example, but it has a big problem in that it can only be wiped and rebuilt a limited number of times. Flash memory is being replaced with NVRAM chips, which have a longer lifespan.

**STREAMING ANALYTICS:**

Analysis of massive amounts of current and "in motion" data using continuous queries known as event streams is known as streaming analytics, or event stream processing. When a particular event, such as a financial transaction, equipment failure, a social media post, or a website visit, occurs as a direct consequence of a specified action, these streams are activated. IoT, transactional and cloud-based data sources as well as online and mobile device interactions and machine sensors may all contribute to the data collection effort. Organizations can derive business value from data in motion using streaming analytics solutions in the same way that they do with data that is at rest. A wide variety of sectors benefit from real-time streaming analytics because they can identify both opportunities and threats.

**EDGE COMPUTING:**

Edge computing is a distributed IT architecture in which client data is handled at the network's edge, as near as feasible to the point of origin.

Nowadays, businesses can't function without data, which gives them access to crucial insights and allows them to exert real-time control over key business processes and activities. Big data may be regularly acquired from sensors and IoT devices working in real time from distant places in hostile situations practically any place in the globe today's enterprises immersed in an ocean of data.

However, this virtual deluge of information is also altering the way corporations use computers. Because of the inability of the old computer paradigm to handle the ever-increasing streams of real-world data, it has been abandoned. Such attempts may be hampered by a variety of factors, including a lack of available bandwidth, long latency times, and random network outages. Using edge computing architecture, businesses are finding a way to deal with these issues.

For the uninitiated, edge computing is just a method of storing and processing data closer to the point at which it is generated. Instead of sending raw data to a central data center for processing and analysis, this work is done where the data is really created — whether that's a retail shop, a manufacturing floor, a vast utility or across a smart city. All of the computing work done on the edge is delivered back to the main data center only to be reviewed and interacted with by humans in the form of real-time business insights or other actionable solutions.

As a result, edge computing is transforming both IT and business computing. Look at edge computing from many angles, including how it works, how the cloud affects it, and how edge use cases, tradeoffs, and implementation issues affect it.

**ARTIFICIAL INTELLIGENCE:**

A field of computer science known as artificial intelligence (AI), often known as machine intelligence, is concerned with developing and managing technology that can learn to make choices and carry out activities on behalf of a human person.

Artificial intelligence (AI) is not a single technology. It is an umbrella word that refers to any form of software or hardware component that enables machine learning, computer vision, natural language understanding (NLU), and natural language processing (NLP).

Today's artificial intelligence makes use of regular CMOS hardware and the same fundamental computational operations that power traditional software. Future generations of artificial intelligence are predicted to inspire new sorts of brain-inspired circuits and architectures that can make data-driven judgments quicker and more precisely than a human person.

**APACHE SPARK:**

Data processing engine Apache Spark (Spark) is a free and open-source tool for huge data collections. To meet the needs of Big Data, it is intended to provide the computing speed, scalability, and programmability necessary for streaming data, graphs data and machine learning.

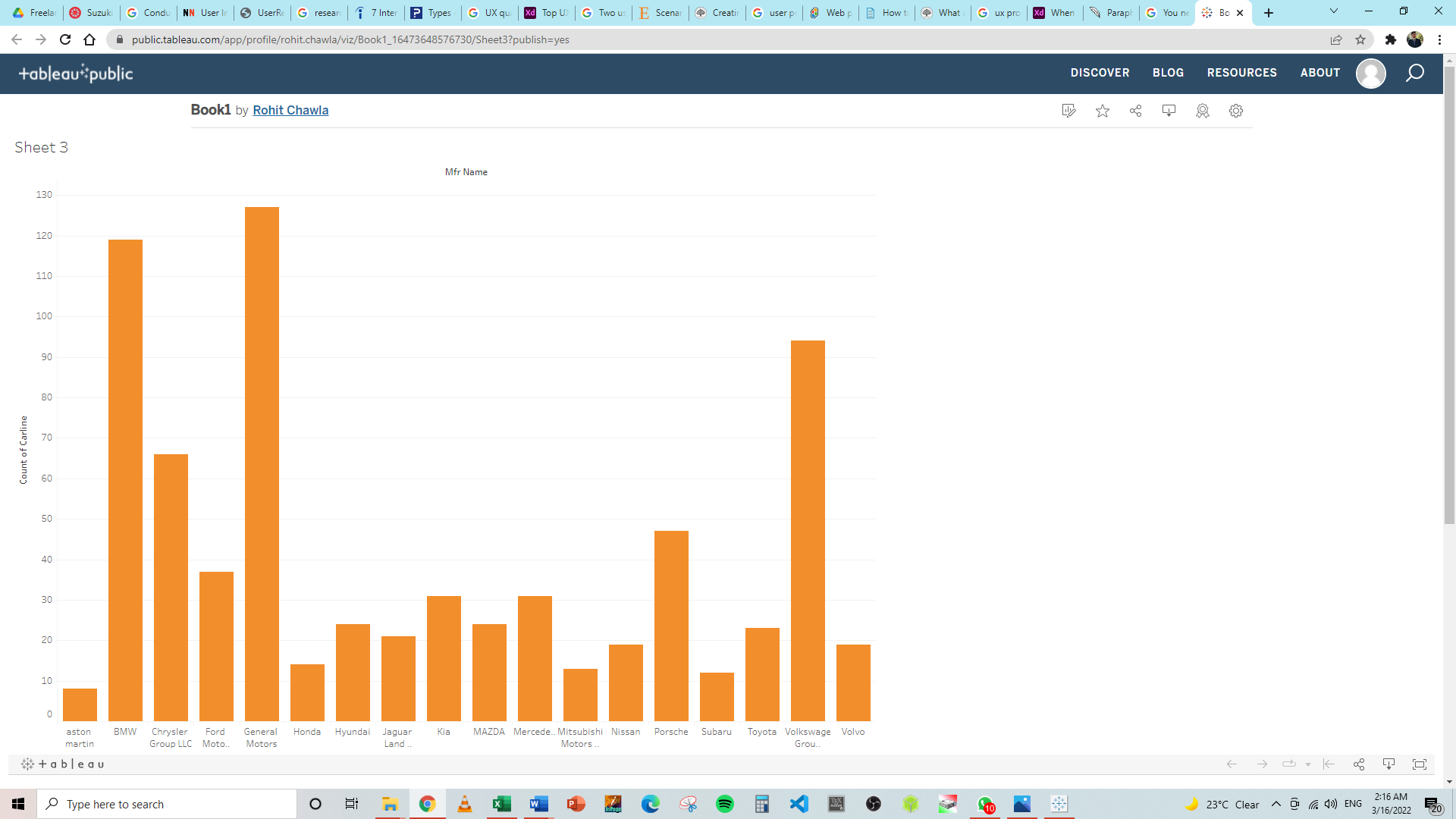
Spark's analytics engine can analyze data up to a hundred times quicker than competing solutions. With built-in parallelism and fault tolerance, it can grow over massive computer clusters. Scala, Java, Python, and R are just some of the prominent programming languages used by data analysts and data scientists.

Often, Spark is compared to Apache Hadoop's native data processing component, MapReduce. Unlike MapReduce, Spark processes and stores data in memory for following stages without writing to or reading from disc, resulting in much quicker processing times.

**TASK 2:**

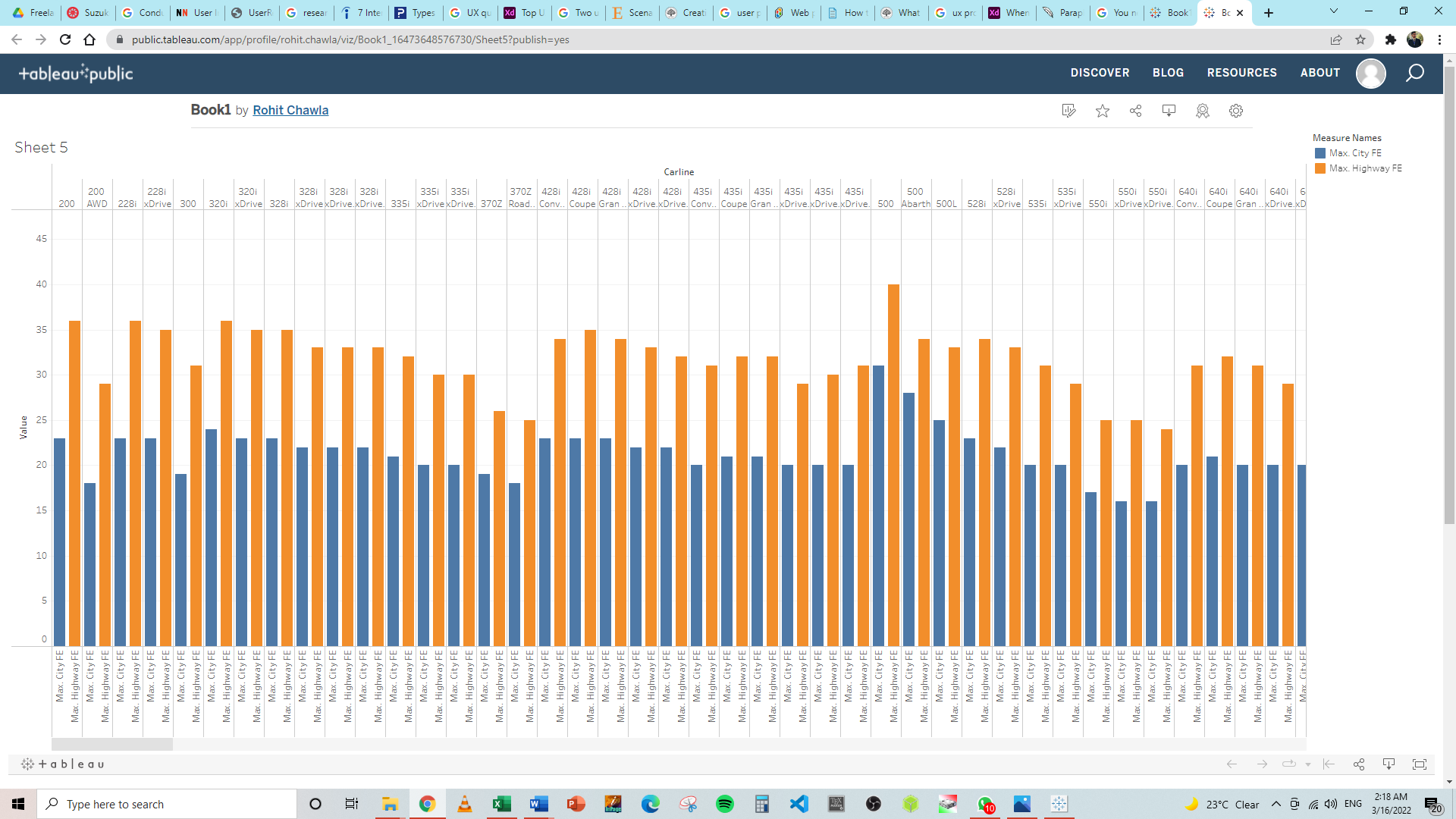
1. Find out which car manufacturer produces the lowest quantity of models.

e.g. BMW 3 Series and BMW 5 Series are different models**.**



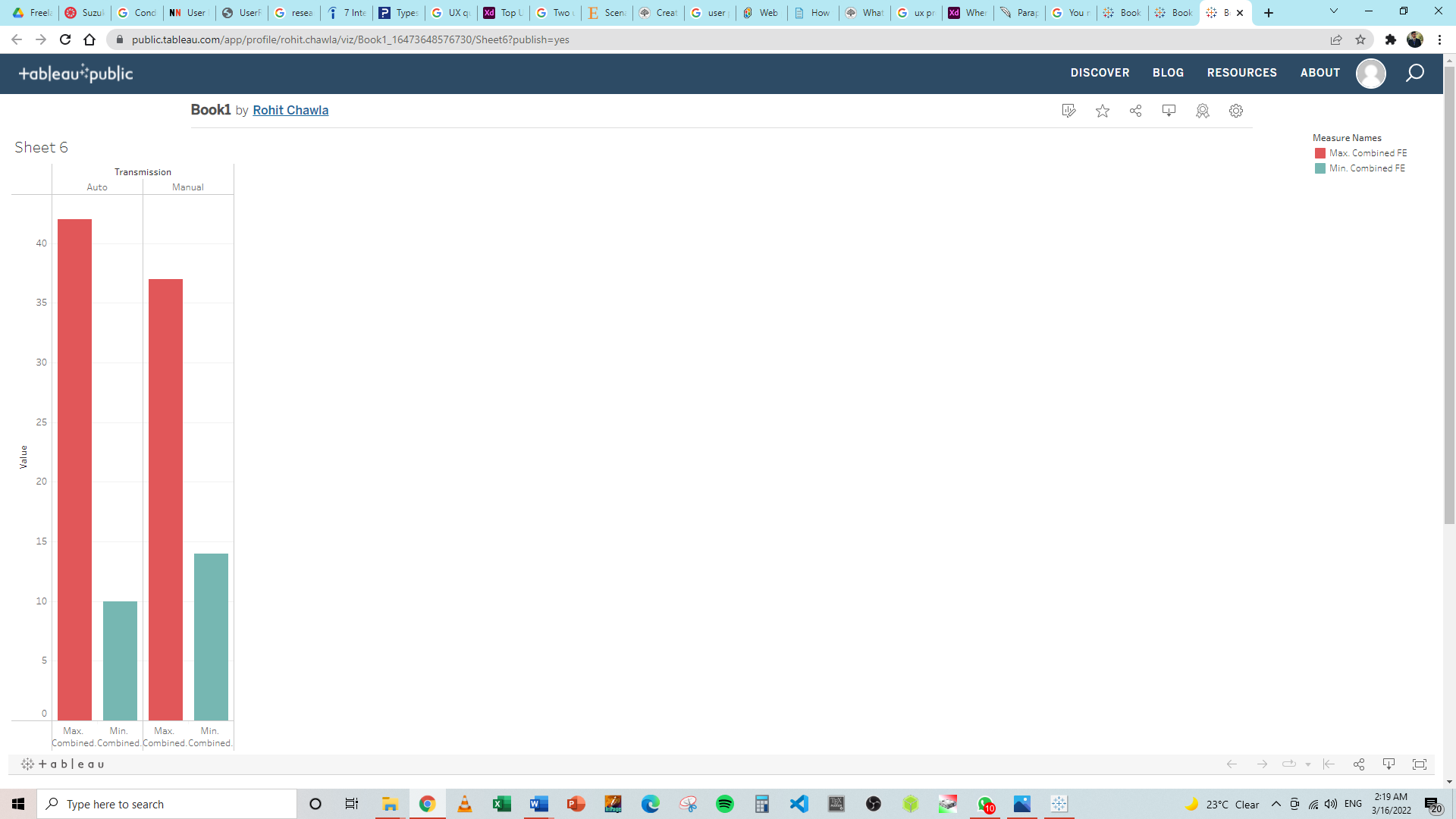
Above is the snippet of the data extracted through tableau from the provided car data where we can conclude that Aston Martin is the least produced or manufactured car because of the luxurious and its unique design and it lies in the super cars and sports car category that’s why it has the limited market and limited customers which demand this type of vehicle. Thus, having a sports car could be a dream but everybody couldn’t afford it unlike other brands listed above in the graphical representation.

1. Find out the highest average fuel economy for urban and motorway driving from the given data set.



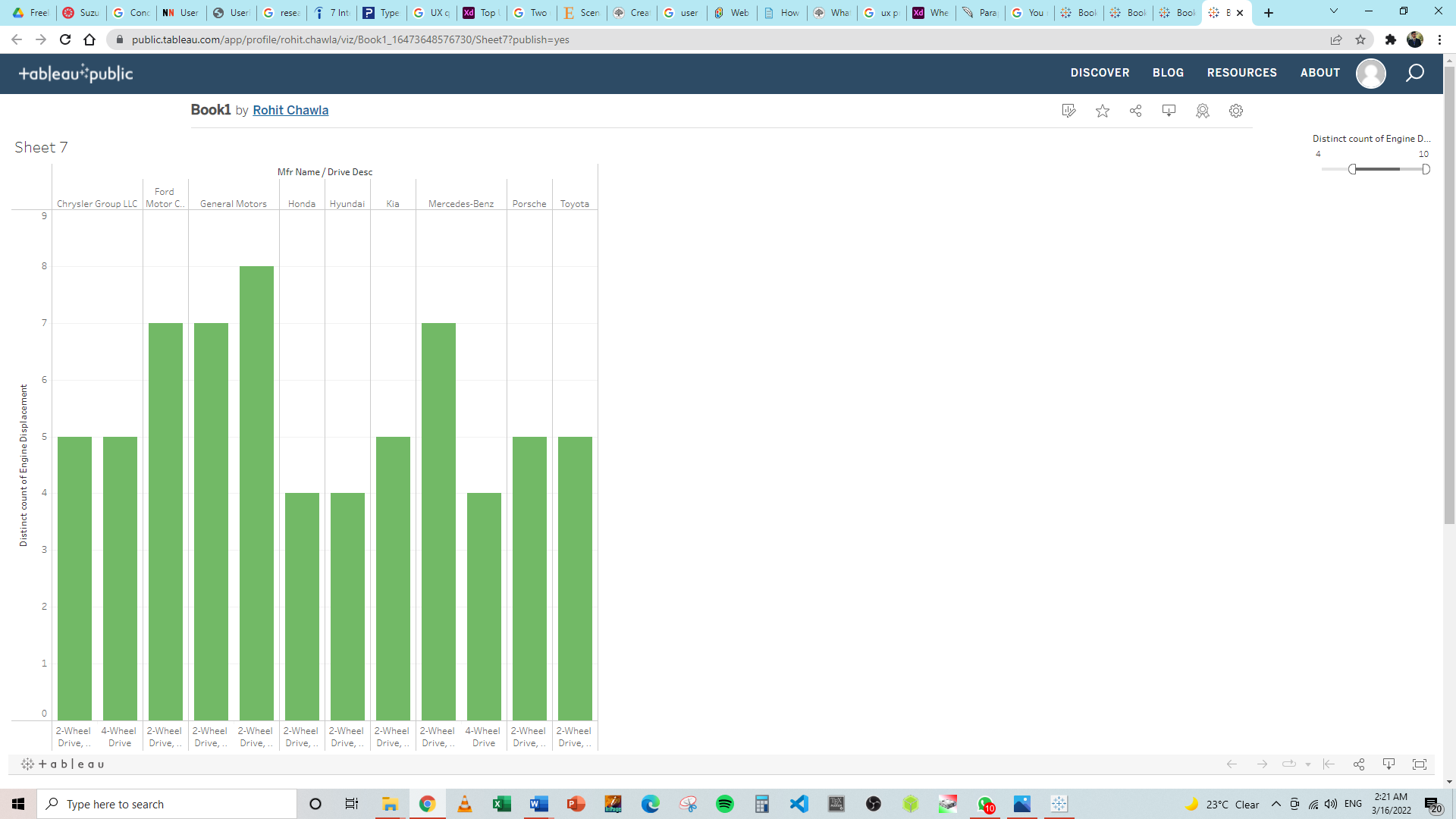
The highest average fuel economy for urban and motorway driving from the given data set extracted and visually represented by tableau is the Fiat 500 car which got the high fuel efficient by taking the constraints of carline all models and the comparison of all model’s mileage mentioned in the car data provided.

1. Find out the high and low average fuel economy from all transmission types.



In the car dataset provided the combined fuel mileage of each car was mentioned according to the auto and manual transmissions thus the tableau represented visually and results found that the auto transmission cars got the high and low fuel average economy as they are run efficiently by the auto transmission rather than manual which some times gives un predictable mileages depending upon the driving of car.

1. Find out which car manufacturers have 4WD (4 wheel drive) and 2WD (2 wheel drive) models and whose engine power is more than 3.5.



According to the results provided by tableau manipulation by the provided car data now we can conclude that Chrysler, Ford, General Motors, Honda, Hyundai, Kia, Mercedes-Benz, Porsche and Toyota are the brands which provide 2WD and 4WD options whereas the cars which have engine power more than 3.5 includes Chrysler, Ford, General Motors, Honda, Hyundai, Kia, Mercedes-Benz, Porsche and Toyota.

1. Critically evaluate the strengths and weaknesses of data analytics, using Tableau and other recommended tools (up to 3 tools) which are used for data analytics.

The strength and weaknesses of data analytics are stated below:  
  
**Strength**

* **An company may make better judgments thanks to the use of data analytics**

Organizational choices are often based on intuition rather than facts or statistics. A lack of quality data to aid in decision making may be a contributing factor. Analytics may help CEOs make better choices by translating the data they have into useful information. Since incorrect judgments have a detrimental influence on many areas, such as corporate development and profitability, this might be an important source of competitive advantage.

* **Boost the productivity of the workforce**

With the use of analytic tools, enormous volumes of data may be swiftly analyzed to assist accomplish particular corporate objectives. Sharing insights from analytics data with staff creates a culture of efficiency and cooperation. A company's shortcomings and potential areas for development become apparent, allowing for the implementation of corrective measures that will boost overall production.

* **Your customers' behaviour changes are constantly monitored via analytics.**

Customers now have a wide range of options to choose from. Organizations may quickly go into a downward spiral if they aren't responsive to the wants and needs of their customers. In this digital age, customers are always being exposed to fresh information, which causes them to often alter their thoughts. It is almost difficult for enterprises to make sense of all the changes in customer perception data without the power of analytics, thanks to the massive amounts of consumer data. Using analytics, you may learn how and whether your target market's thinking has changed. To stay ahead of the competition, organizations need to be aware of how their customers' buying habits are changing.

* **Product and service personalization**

There is no longer a conventional set of items and services that a corporation may offer to its clients. Customers are looking for goods and services that may be tailored to their specific requirements. To better understand what the consumer likes, organizations use analytics to monitor what they like and then display them what they want depending on their choices. Companies' data collecting and analytics make it possible for us to view just the content we want on social media, for example. Data analytics may assist in providing consumers with personalized services based on their specific needs.

* **The quality of goods and services must be improved**

By recognizing and rectifying mistakes and minimizing non-value-added chores, data analytics may assist improve the user experience. Data from self-learning systems, for example, may be used by these systems to better understand how consumers engage with their products and to make modifications that enhance the user experience accordingly. Furthermore, data analytics may assist in the automatic cleaning of data and the improvement of data quality, all of which benefit consumers and companies alike.

**Weaknesses**

* **Team disunity**

Across-team or cross-departmental misalignment It's possible for a small group of executives to analyze data. While these teams' insights are valuable, their influence on organizational measures is limited. Many teams function in “silos”, utilizing just their own methods and without collaborating with others. In order to positively affect the company, data analytics teams should be focused on addressing business-critical issues and correctly communicating their findings to the appropriate personnel and stakeholders.

* **Inability to commit and wait**

Simple to install, yet expensive and slow to pay off. It takes time to set up systems and procedures to gather data, especially if no data is available now. To deploy the solution, the analytics models need to enhance accuracy over time. Business users lose interest and confidence when they don't see instant outcomes. In order to know what is working and what is not, an organization must have a feedback loop and process in place. A lack of a closed loop system might lead senior management to reject analytics altogether.

* **Data poor quality**

Lack of access to quality data is one of the major data analytics challenges. The issue is whether or not businesses currently have access to enough data. The business questions to be addressed must be understood first, before determining what data is necessary to answer them. A historical dataset may not always be appropriate to solve today's problems. The quality of the data gathering may be inadequate even if we have the proper measures. In certain cases, there may be insufficient data to do meaningful analytics. It's true what they say. Unreliable data results in unreliable decisions. So, before data can be utilized efficiently inside businesses, it must be improved in terms of quality.

* **Privacy issues**

This means that organizations whose services people use may access data about their purchases, online transactions, and subscriptions. Some firms may trade datasets with other companies for mutual gain. Some data may be exploited against a person, nation, or society. In order to protect client data, businesses must be careful in collecting it. To secure sensitive data, only data essential for the analysis should be gathered. The loss of client trust may negatively affect an organization's bottom line.

* **Bias & Complexity**

Some firms' analytics tools are more of a black box. A model is created through learning from data. Determining who should be approved for a loan using a neural network model that learns from different instances. However, the company's decision-making process is illogical. These systems may have hidden biases in their conclusions that are not easily apparent if corporations are not attentive and utilize low quality data to train the model.

**REFERENCES:**

[**https://snapstack.cz/what-is-edge-computing-and-what-are-its-advantages/**](https://snapstack.cz/what-is-edge-computing-and-what-are-its-advantages/)

[**https://www.linkedin.com/pulse/what-artificial-intelligence-really-margaret-rouse**](https://www.linkedin.com/pulse/what-artificial-intelligence-really-margaret-rouse)

[**https://www.techtarget.com/searchdatamanagement/definition/data-lake**](https://www.techtarget.com/searchdatamanagement/definition/data-lake)

[**https://databricks.com/glossary/streaming-analytics**](https://databricks.com/glossary/streaming-analytics)

[**https://aws.amazon.com/nosql/in-memory/**](https://aws.amazon.com/nosql/in-memory/)