

**ITCS 6100-Big Data Analytics**

**Project Deliverable 1**

**New York City Taxi and Limousine Commission (TLC) Trip Record Data**

**Submitted By Instructor**

Srushti Khot Prof. Pamela Thompson

Sanket Revadigar

Nidhi Gowri Srinath

Sharat Sindoor

Sudarshan Ramesh

1. **Team**
2. **Members**
3. Srushti Khot
4. Sanket Revadigar
5. Nidhi Gowri Srinath
6. Sharat Sindoor
7. Sudarshan Ramesh
8. **Communication plan to include project artifact repository.**

Shortlisted few topics from artifact repository based on the information each of them provided. Each one of us came up with 2 topics and finalized a topic we thought would bring different types of analysis.

1. **Selection of data to analyze from the Open Data Registry for Amazon Web Services** [**https://registry.opendata.aws/**](https://registry.opendata.aws/)

This registry exists to assist people in discovering and sharing datasets made available through AWS resources. There are over 300 datasets to choose from, and the dataset below was chosen from a pool of ten by all members.

**Choice of Dataset-** New York City Taxi and Limousine Commission (TLC) Trip Record Data

1. **Business Problem or Opportunity, Domain Knowledge**

New York City, as one of the most populated cities in the United States, has millions of cabs journeys every month. The goal of this project is to gain a better understanding of the trip record data and make some ideas that might help taxi drivers earn more money.By studying user trends, an organization's profit may be increased in a variety of ways. As a result, we'll need certain approaches to evaluate user activity and develop patterns as a result.

The yellow and green taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts. The data used in the attached datasets were collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP).

**Dataset Structure**

VendorID int64

tpep\_pickup\_datetime datetime64[ns]

tpep\_dropoff\_datetime datetime64[ns]

passenger\_count int64

trip\_distance float64

RatecodeID int64

store\_and\_fwd\_flag object

PULocationID int64

DOLocationID int64

payment\_type int64

fare\_amount float64

extra float64

mta\_tax float64

tip\_amount float64

tolls\_amount float64

improvement\_surcharge float64

total\_amount float64

trip\_dur\_secs int64

pickup\_date object

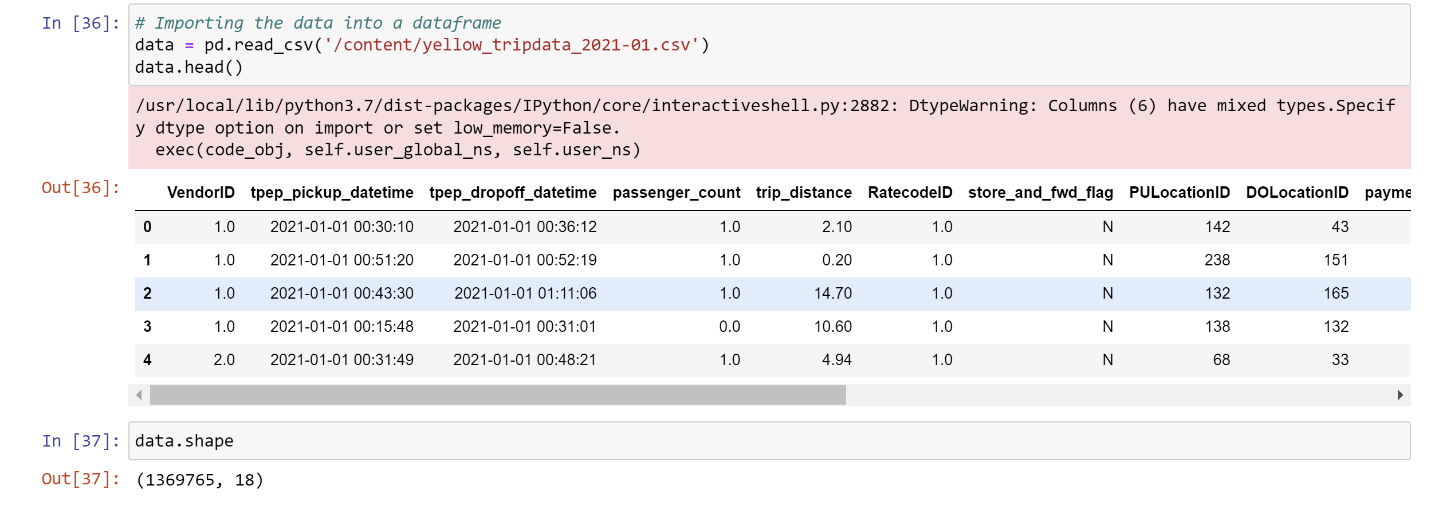
1. **Research Objectives and Question(s)**

Through our project we intend to answer and visualize the following questions:

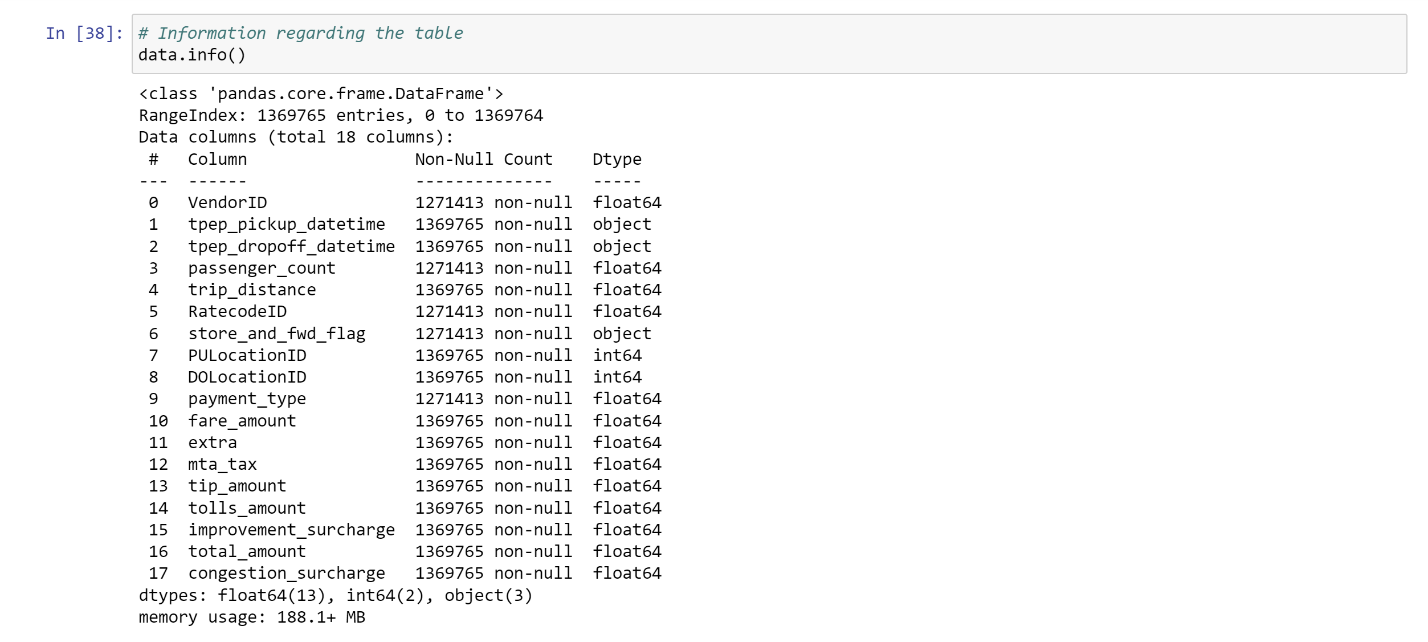
* What are the general trends of the pickup and drop-off locations of taxis in New York City?
* What timings experience the most surge in taxi bookings?
* What are the fare trends during these surge timings?
* What are the preferred methods of payment for taxi fares?
* Is there a correlation of tips with various factors such as trip distance, location, taxi booking timings, etc.?
* What is the contrast of booking frequency on weekdays vs. weekends and peak hours vs. non-peak hours?

1. **Data Understanding**
2. **Exploratory Data Analysis**

Data is expanding at an exponential rate, is arriving from new sources, is becoming more diversified, and must be safely accessed and analyzed by a rising number of applications and individuals. Because of the scale, complexity, and variety of data sources, the same technologies and tactics that worked in the past no longer work. When developing analytical models, it is advisable to start with tried-and-true open datasets from the issue domain under consideration. This allows us to configure our data analytics pipeline, select relevant models and analytical methodologies, assess the outcomes, and communicate with the open data community before applying them to our own data. Such open datasets may be found on AWS's Registry of Open Data.

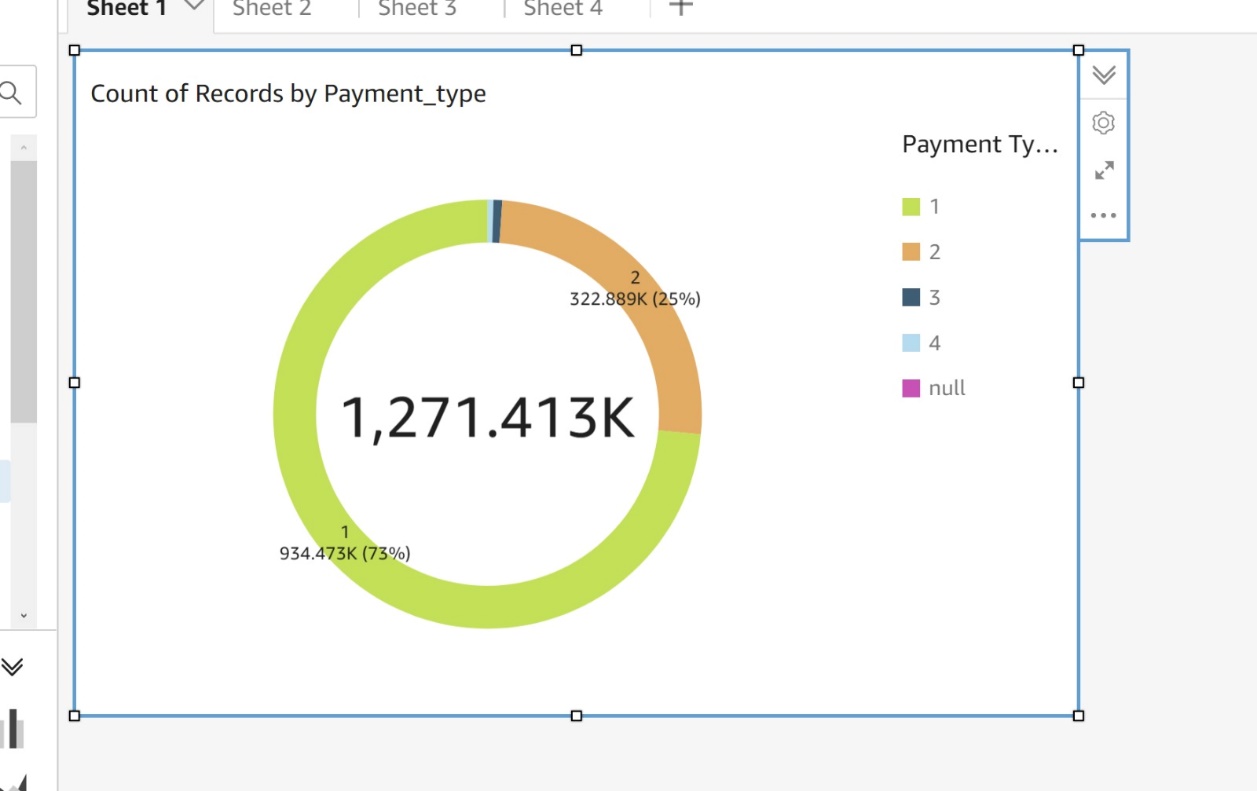


* To get insights, a dataset of January 2021 yellow taxis is employed.
* The dataset contains 1369765 records spread into 18 columns.

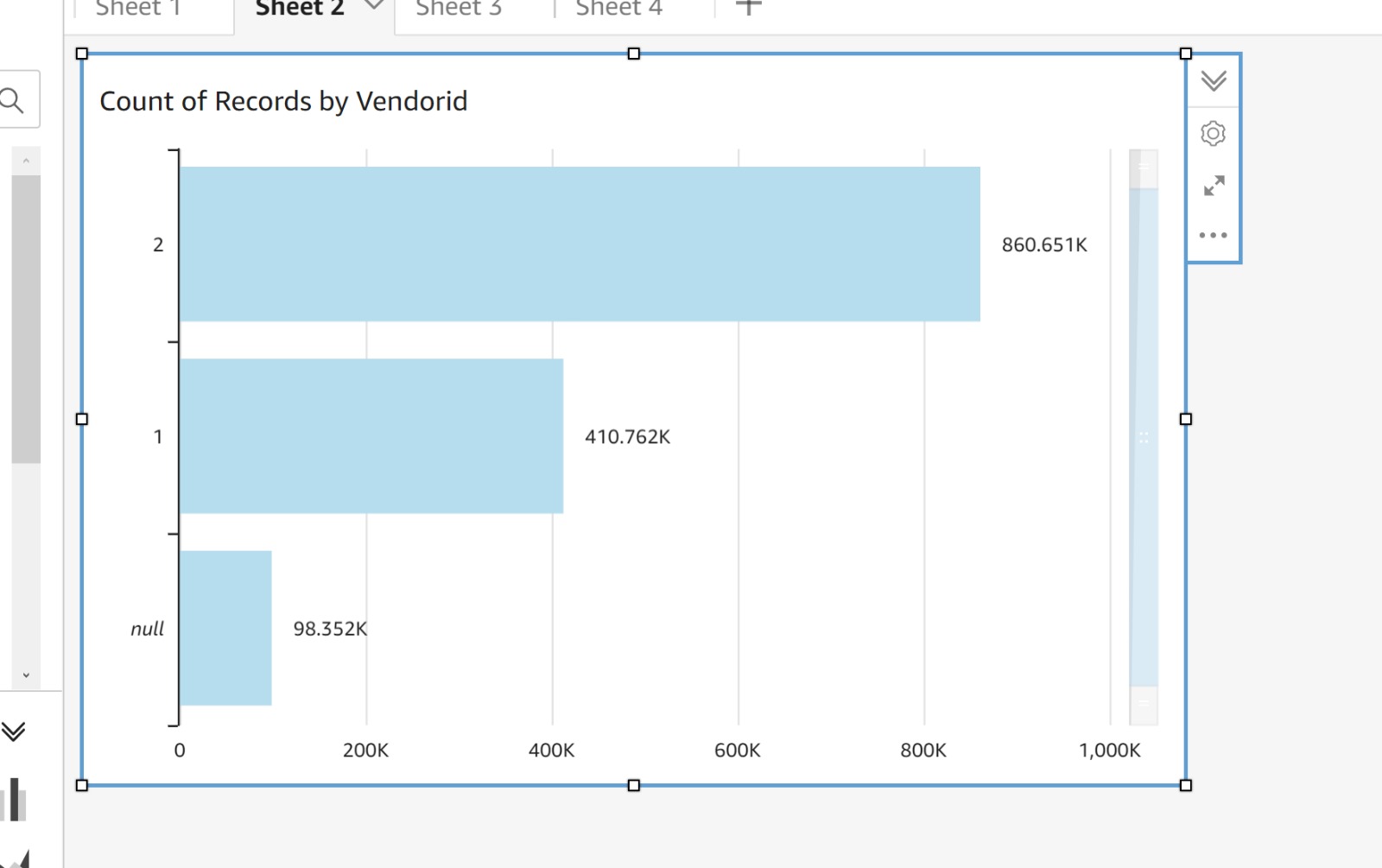


* The above screenshot represents all the column names, data type, and count of non-null entries present in the dataset.

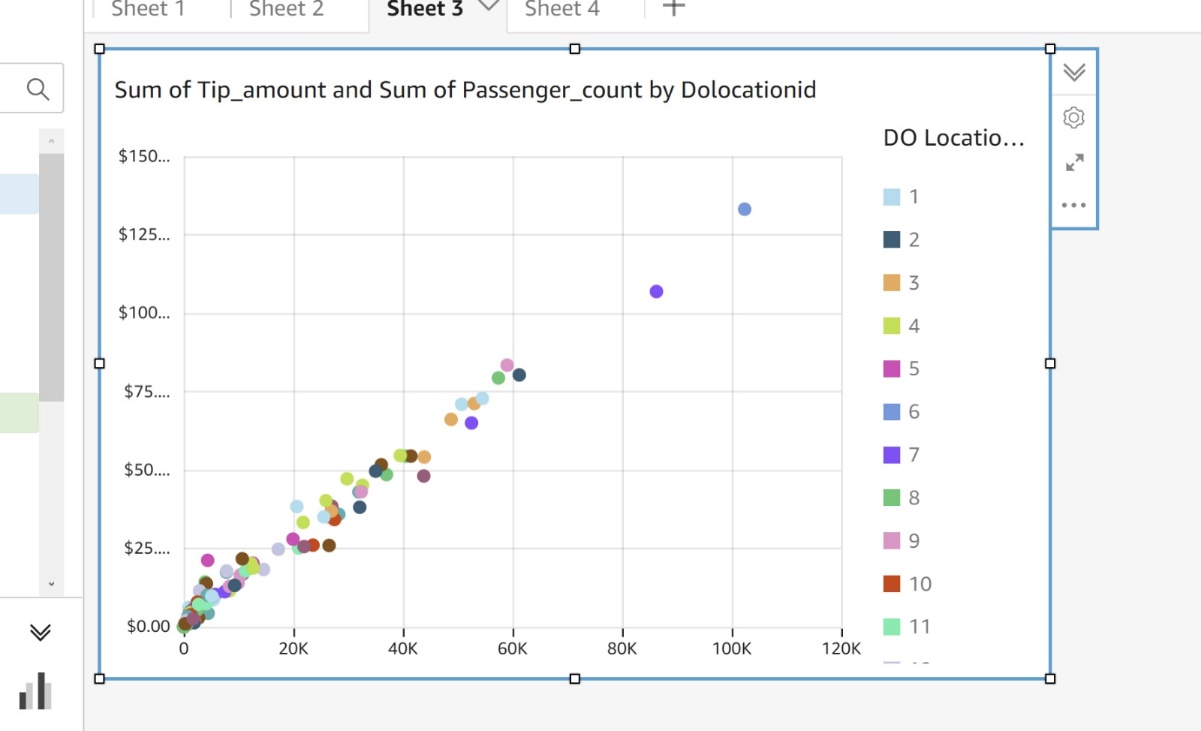
1. **Dashboard**

****

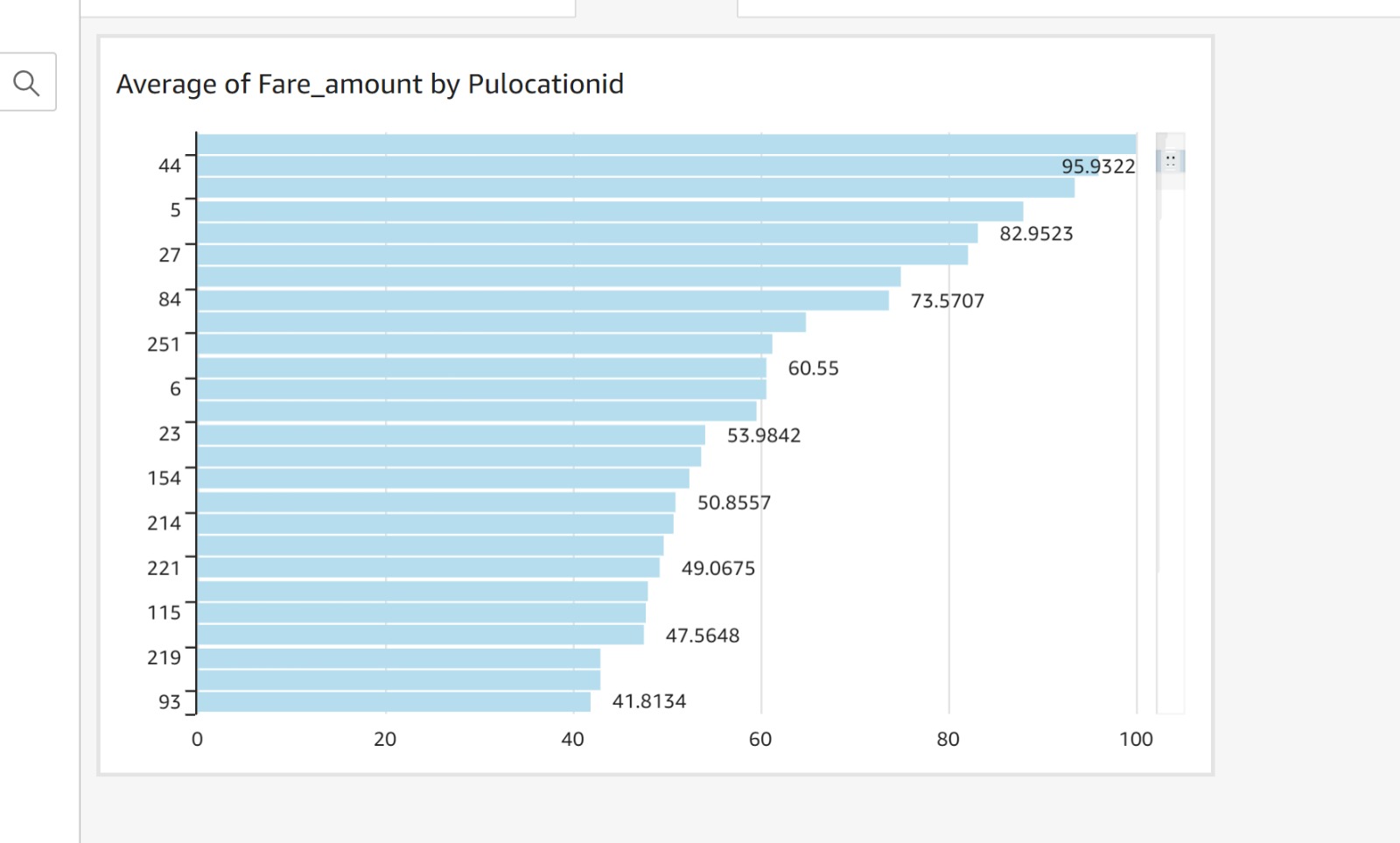
* As we can see from the dashboard we have 1271.413k records of payments by passengers



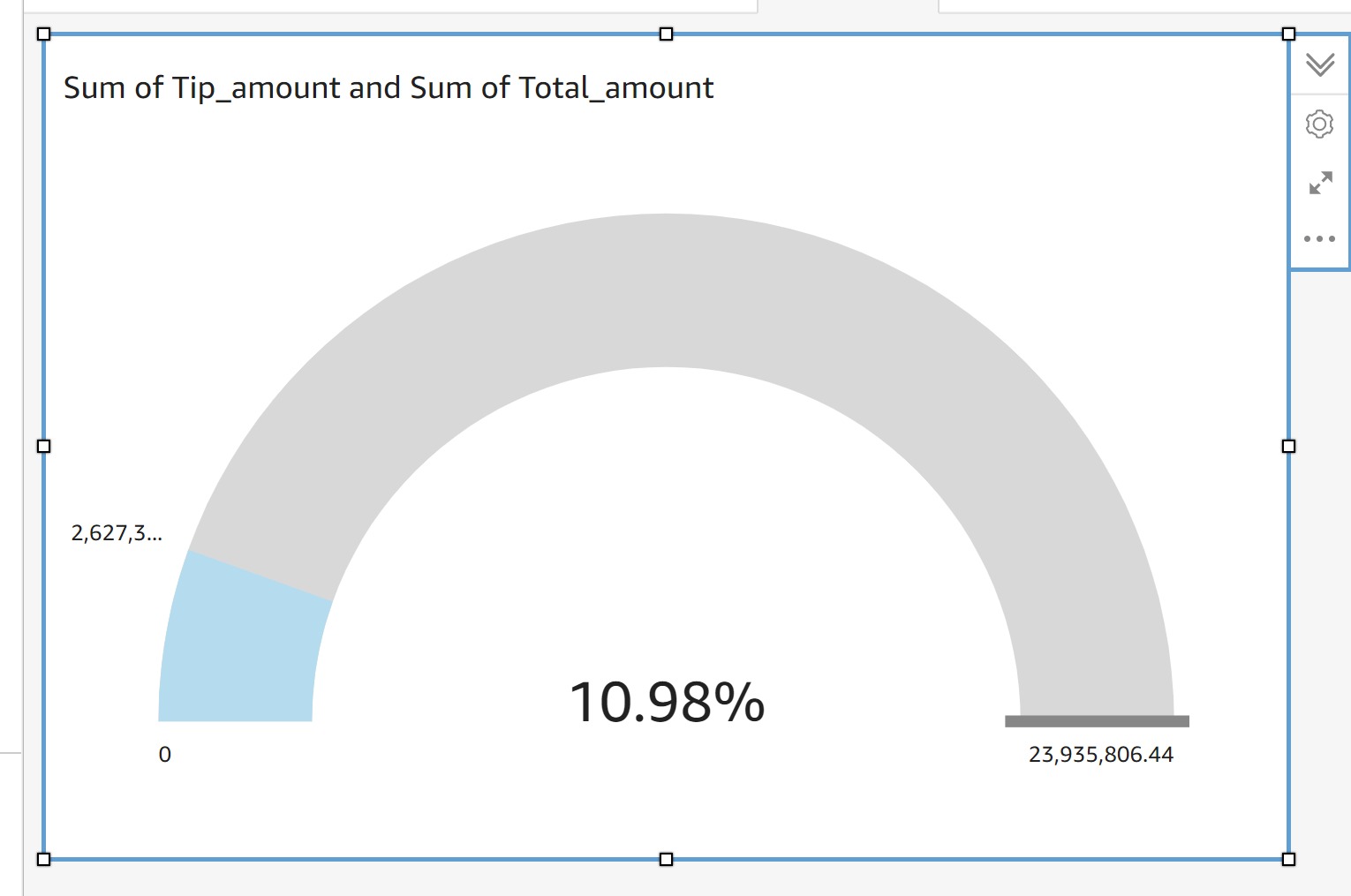
* Count of records from different vendors (VendorId - 1 and 2). Null represents count of rows that have missing values. Vendor 2 has highest customers.



* Graph showing relation between Sum of Tip amount and Sum of passenger count filtered by drop of location.



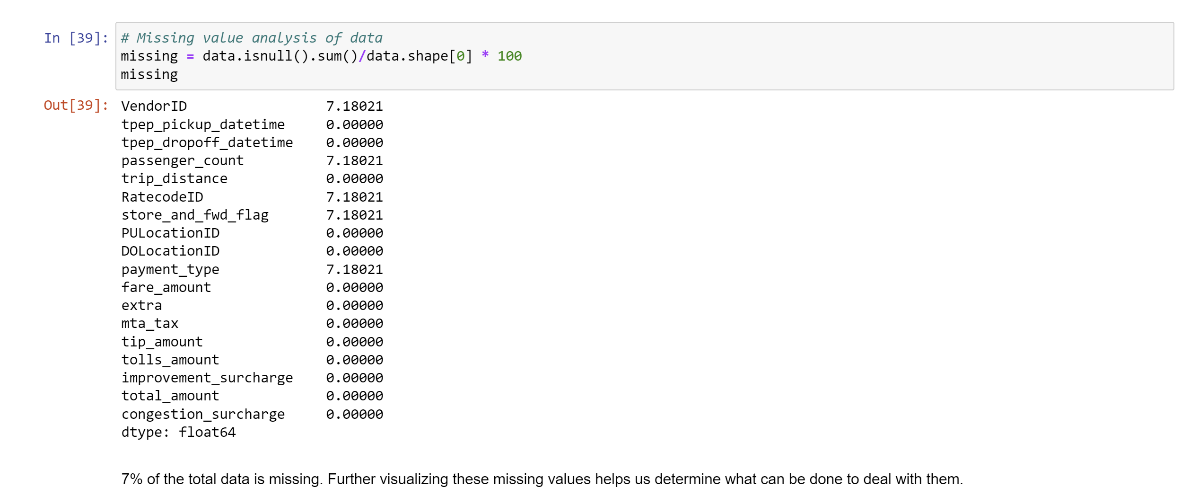
* Average fare of the trip based on the pickup location. The location id 44 has the highest fare with close to 100$.



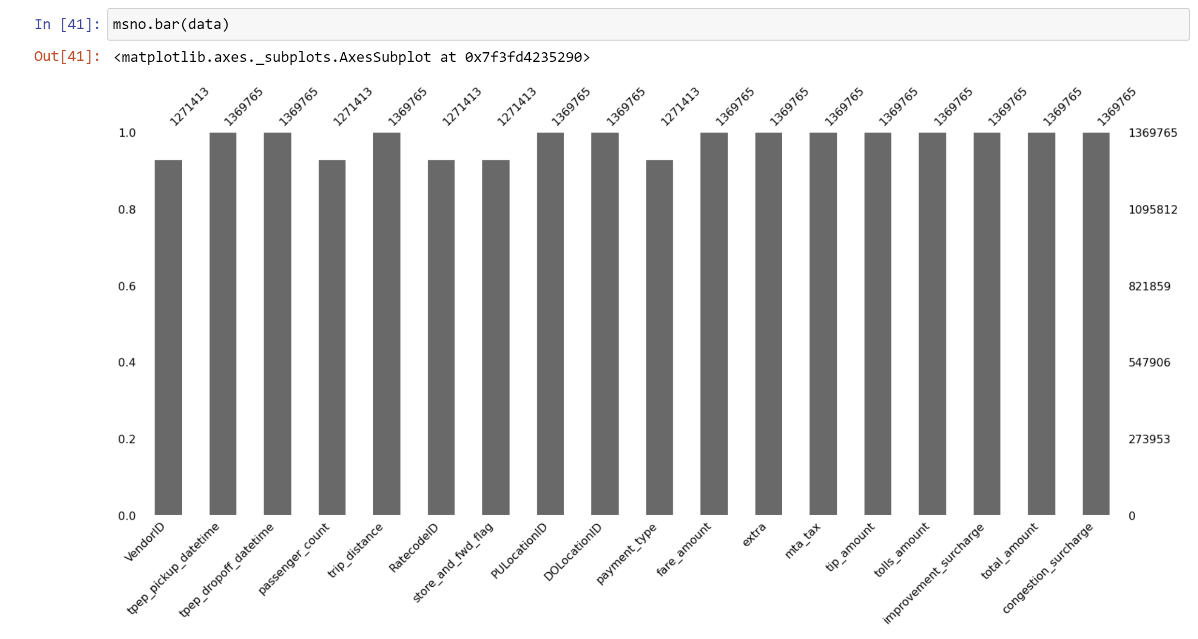
* Average tip amount for each ride, which is around 11%

1. **Data Preparation**

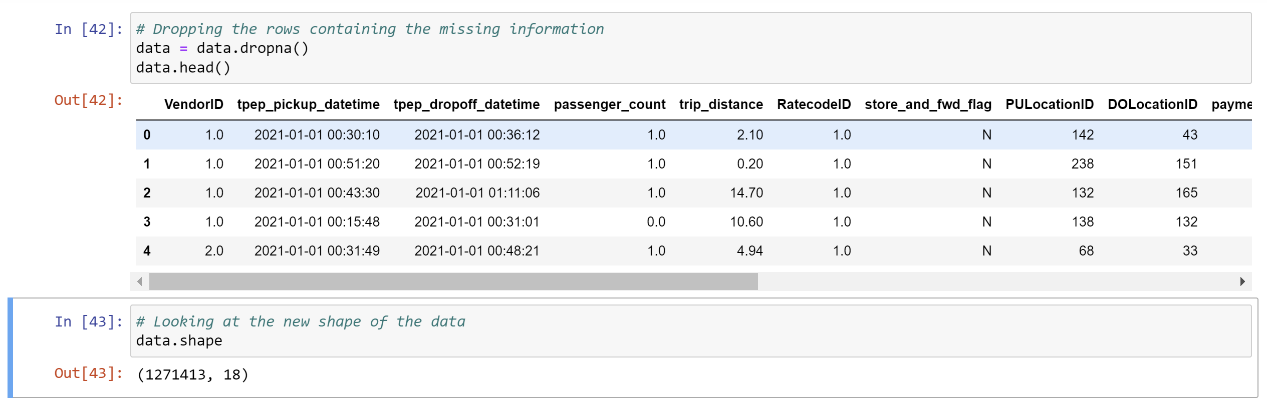
The act of obtaining, integrating, formatting, and organizing data so that it may be utilized in business intelligence (BI), analytics, and data visualization applications is known as data preparation. Data preparation includes data pretreatment, profiling, cleaning, validation, and transformation; it also frequently entails bringing together data from many internal and external systems. Informally, data preparation is referred to as data prep. It is also referred to as data wrangling, while some practitioners use the phrase in a narrower meaning to refer to cleaning, organizing, and transforming data; this usage separates data wrangling from the data preparation step.



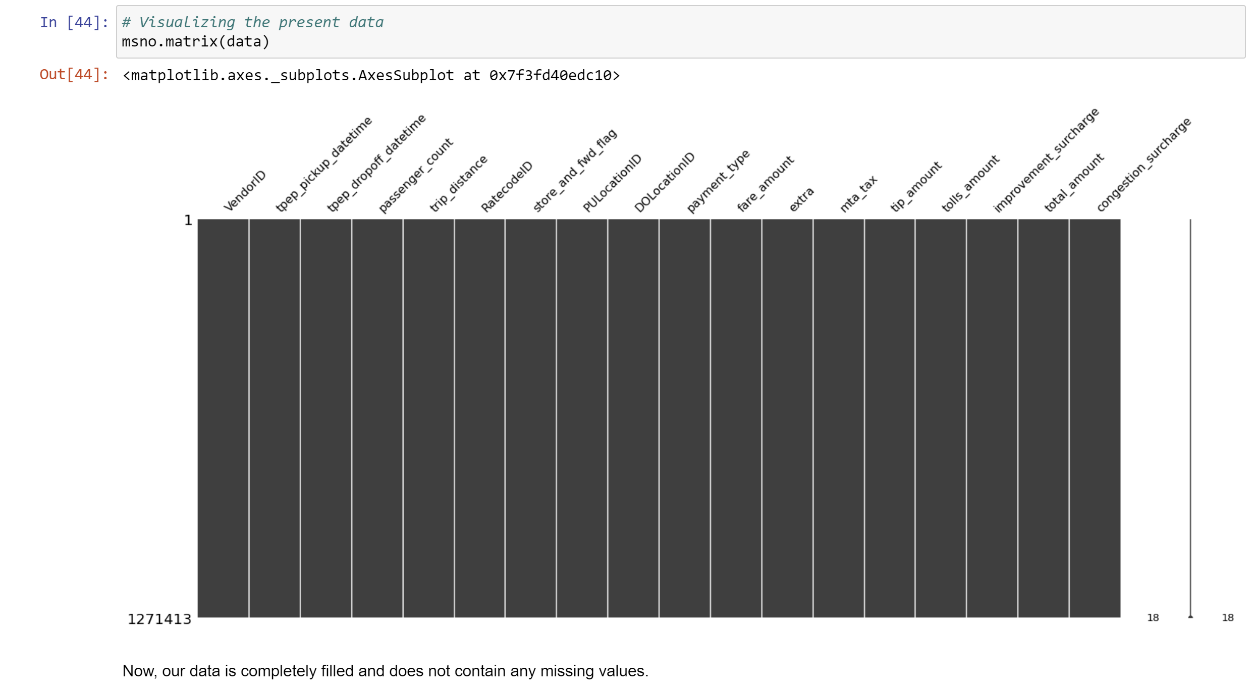
* This screenshot shows percentage of missing values in each column.So our task is to eliminate all the missing values.



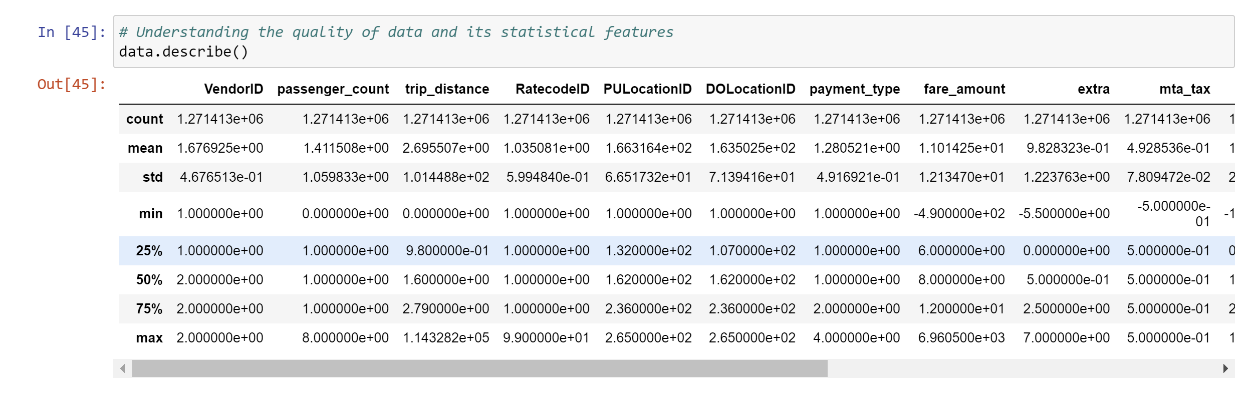
* A bar chart to show total no of records in each column.
* Here, we see that the missing values lie in 5 columns.
* Vendor ID, Passenger Count, Rate Code ID, Store and Forward Flag, Payment Type are the columns with missing values.



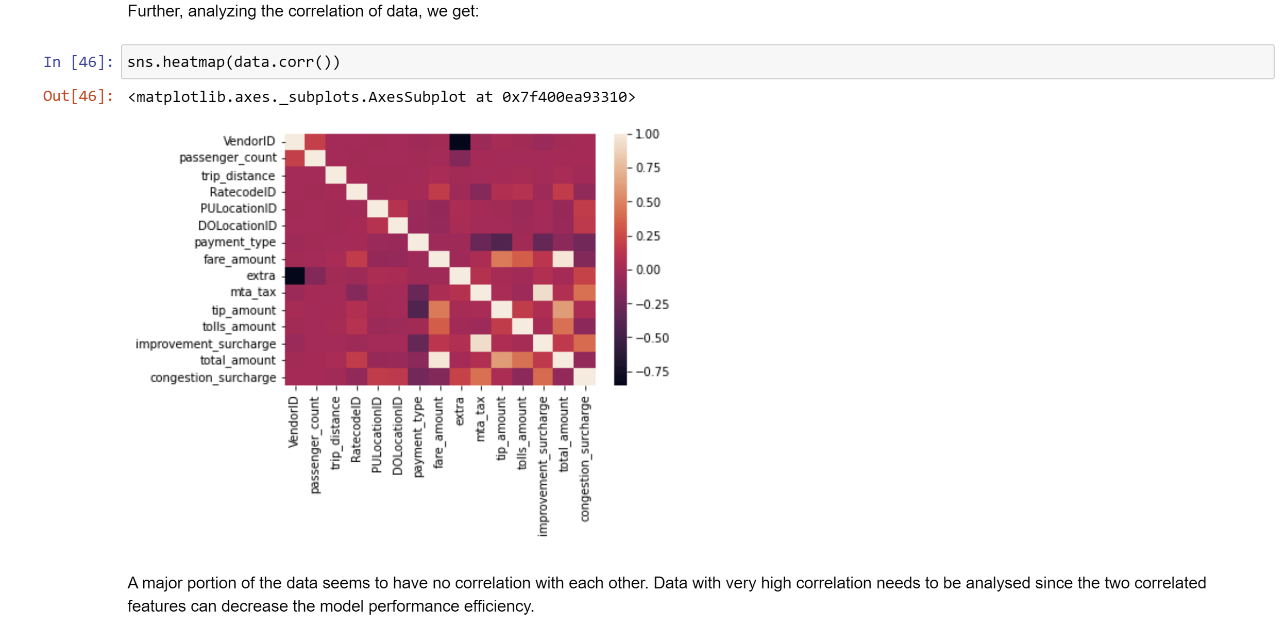
* After removing the rows with missing values, the total number of records in 18 columns will be 1271413.



* We can now observe that the number of records in each column is the same.



* Understanding the quality of data and its statistical features, with the help of data frame describe function.



* The data does not appear to be related in any way. Data with a high degree of correlation must be examined since the two associated properties can reduce the model's performance efficiency.

1. **References**

* GitHub repository - <https://github.com/SanketRevadigar/ITCS-6100---BDA>
* <https://registry.opendata.aws/>
* <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>
* <https://aws.amazon.com/blogs/machine-learning/machine-learning-on-distributed-dask-using-amazon-sagemaker-and-aws-fargate/>
* <https://github.com/aws-samples/cloud-experiments/tree/master/experiments/notebooks/exploring-data>
* <https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation#:~:text=Data%20preparation%20is%20the%20process,analytics%20and%20data%20visualization%20applications>.