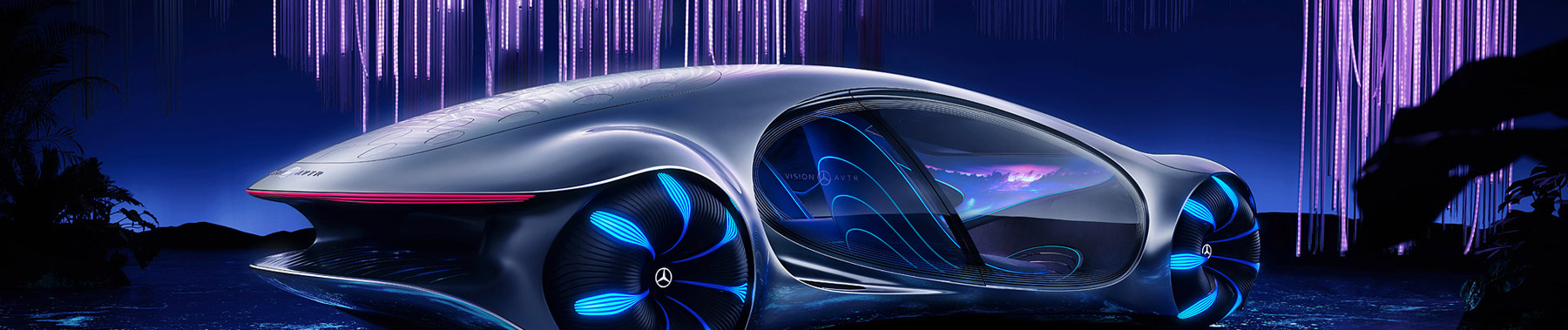
Big-Data Architecture and Governance

Used Cars Price Analysis

Us (United States) Used Cars Dataset

Group 4: Surbhi Wagh, Rohit Singh, Sumit Patil



**Project Overview:**

Due to increasing demand in cars by the people for commuting in US there are group of people who are even interested in purchasing used cars for their use.

In this case it's necessary to analyze the condition brand type model and so on for the car and check the price of cars as per purchaser needs.

This project will help to get an idea of the price of the used cars in United states

**Objective and Tools used:**

**Objective**:

The goal of this project is to find patterns in the data and build a dashboard that help a customer to get an idea regarding the price of the used car according to his requirement and colors and as per his location to get idea about what's the used card available around his area.

**Tools Used:**

Data Extraction, Profiling, Cleaning, Visualization: Python(Jupyter notebook)

Graph Creation: Arrow. App

Graph Database Management System: Neo4j

Project, Resource and Risk Management: Velero ETP

**About this file**

|  |  |
| --- | --- |
| vin: | Vehicle Identification Number is a unique encoded string for every vehicle. |
| back\_legroom: | Legroom in the rear seat. |
| bed: | Category of bed size(open cargo area) in pickup truck. Null usually means the vehicle isn't a pickup truck |
| bed\_height: | Height of bed in inches |
| bed\_length: | Length of Bed in inches |
| body\_type : | Body Type of the vehicle |
| cabin: | Category of cabin size |
| city : | City where the car is listed |
| city\_fuel\_economy: | Fuel economy in city traffic in km per litre |
| combine\_fuel\_economy : | Combined fuel economy is a weighted average of City and Highway fuel economy in km per litre |
| daysonmarket: | Days since the vehicle was first listed on the website |
| dealer\_zip: | Zipcode of the dealer |
| description : | Vehicle description on the vehicle's listing page |
| engine\_cylinders : | Engine configuration. |
| engine\_displacement : | Measure of the cylinder volume |
| engine\_type : | Type of Engine |
| exterior\_color : | Exterior color of the vehicle |
| fleet : | Whether the vehicle was previously part of a fleet |
| frame\_damaged : | Whether the vehicle has a damaged frame. |
| franchise\_dealer : | Whether the dealer is a franchise dealer. |
| franchise\_make : | Company that owns the franchise |
| front\_legroom : | The legroom in inches for the passenger seat |
| fuel\_tank\_volume : | Fuel tank's filling capacity in gallons |
| fuel\_type : | Dominant type of fuel ingested by the vehicle. |
| has\_accidents : | Whether the vin has any accidents registered |
| height : | Height of the vehicle in inches |
| highway\_fuel\_economy : | Fuel economy in highway traffic in km per litre |
| horsepower : | Horsepower is the power produced by an engine |
| interior\_color : | Interior color of the vehicle |
| isCab : | Whether the vehicle was previously taxi/cab |
| is\_certified : | Whether the vehicle is certified |
| is\_cpo : | Certified vehicles warranty for free repairs flag |
| is\_new : | Is the vehicle launched less than 2 years ago |
| is\_oemcpo : | Pre-owned cars certified by the manufacturer. |
| latitude : | Latitude from the geolocation of the dealership |
| length : | Length of the vehicle in inches |
| listed\_date : | Listing date of the Vehicle |
| listing\_color : | Dominant color group from the exterior colo |
| listing\_id : | Listing id from the website |
| longitude : | Longitude from the geolocation of the dealership. |
| main\_picture\_url : | Url of the picture of the used Vehicles |
| major\_options : | Major options available for the vehice |
| make\_name : | Brand of the vehicle |
| maximum\_seating: | Maximum number of seats |
| mileage: | Mileage of the car |
| model\_name : | Model name of the cars |
| owner\_count : | No of previous owner |
| power : | power configuration of the car |
| price : | price of the car |
| salvage : | Whether car has been damaged or not |
| savings\_amount : | amount you will save buying car |
| seller\_rating : | Rating of the seller |
| sp\_id : | Seller Id |
| sp\_name : | Seller name |
| theft\_title : | Whether the the vehicle has theft recovery |
| torque : | Torque of the vehicle |
| transmission : | Transmission type which is CVT or automatic |
| transmission\_display : | Displaying the transmission type of the display |
| trimId : | Version of model with configuration |
| trim\_name : | Name of the version |
| vehicle\_damage\_category : | Any damage to the vehicle |
| wheel\_system : | Types of wheel system |
| wheel\_system\_display : | wheel drive systems |
| wheelbase : | wheel base of the vehicle in inches |
| width : | Width of the vehicle in inches |
| year : | The year on which the vehicle was made. |

**Risk & Issues of Project:**

|  |  |  |
| --- | --- | --- |
| Type | Risk/Issue-Description | Mitigation |
| Issue | Improper Environment setup for Data cleansing and data validation i.e., insufficient packages to run python scripts | Make sure all packages are installed and proper environment is set up to run all scripts error for data cleansing/wrangling process |
| Risk | Many fields needed to track cars details has missing data which might be an analytical risk leading to biased visualization | Deriving new columns by understanding existing database attributes to generate reports without errors. |
| Risk | Inconsistent entries in data available for cars classification and thus analysis might be biased eventually which further leads to incomplete data merging or error in the creating graph database | Categorization of the range for available data records is to be done before loading data into the database. |
| Risk | Lack of Knowledge in terms of cars dataset among team members could lead to incorrect analysis and interpretation | Prior research/study of cars and features for understanding dataset |

**Data Profiling Instructions:**

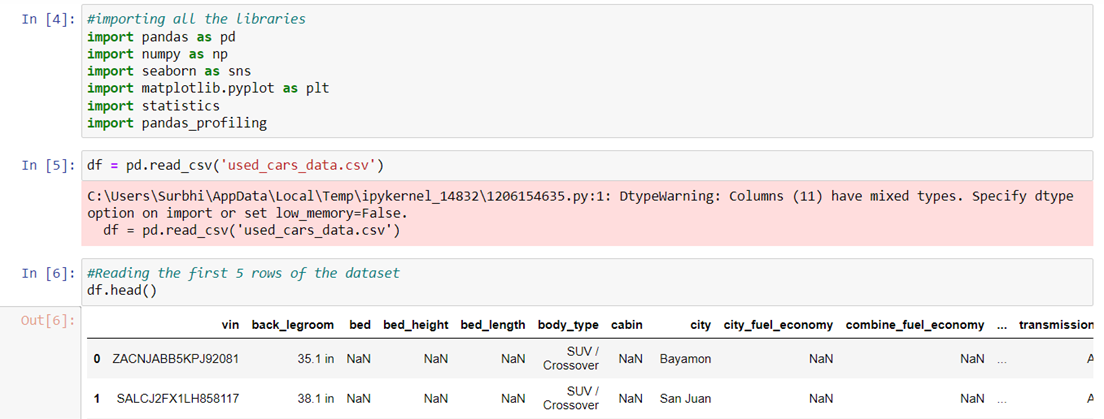
1) Download Python (follow instructions on https://www.python.org/downloads/ based on operating system

2) Install pandas library by running pip install pandas

3) Install pandas profiling to be able to create an html profile report by running the command pip install pandas-profiling

4) Run profiling.py script and output.html file will be created

5) Open output.html on browser to view data profile report



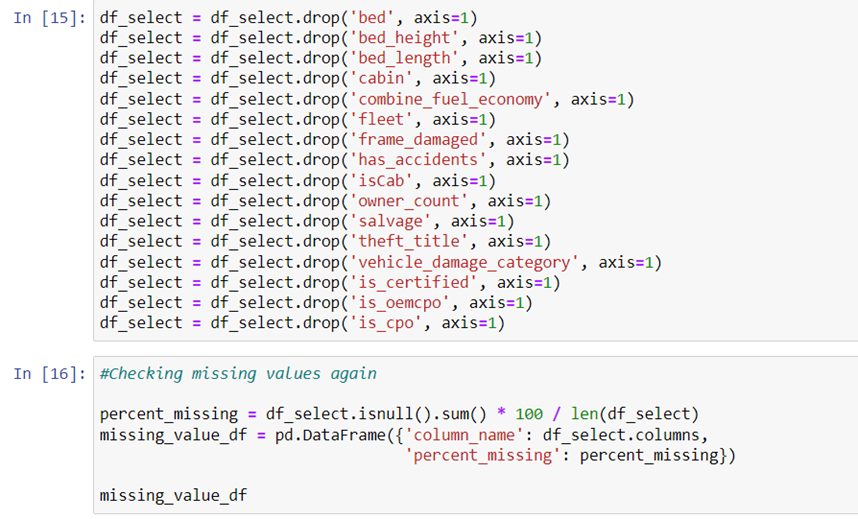
**Data Wrangling/Cleansing Instructions**

1) Make sure pandas library is installed (should have been installed during data profiling)

2) Run exploration.py script

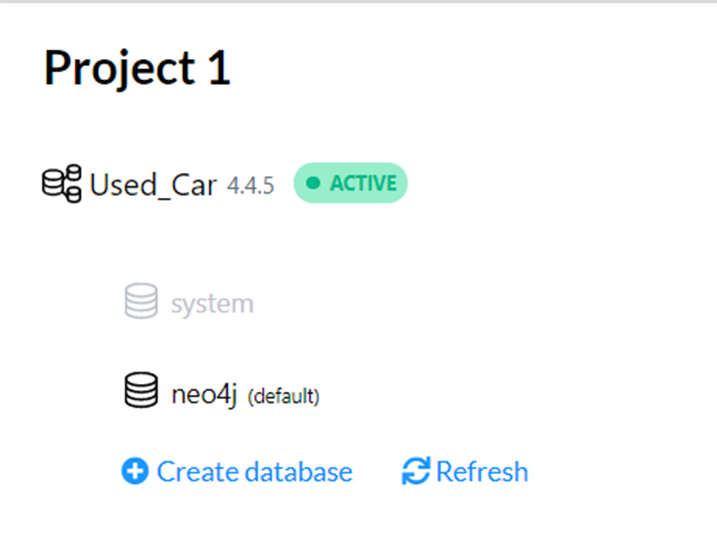
3) Once script is finished running, “old-car-dataset.csv” file will be created

4) Use this new filtered dataset for database/visualization

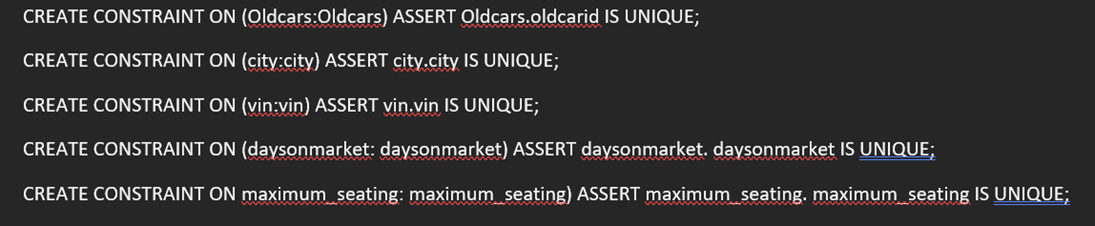


**Neo4j Instruction:**

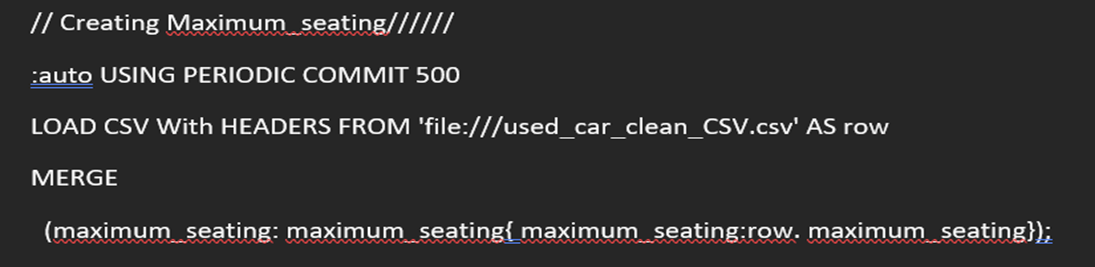
1. Create a new database in Neo4j
2. Copy the CSV file into the dbms import folder of the database created
3. Start the database and open the database



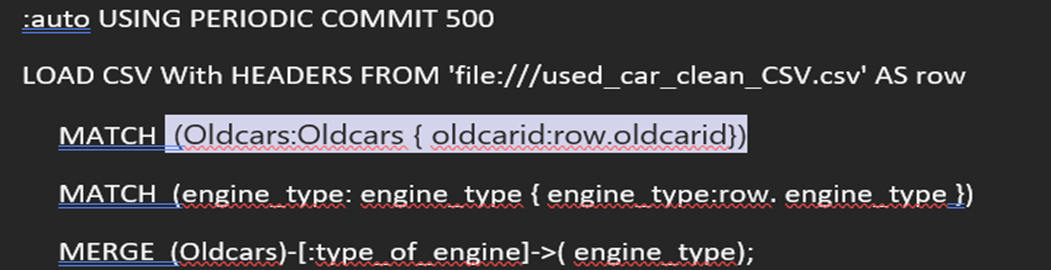
**Constraint Creation:**



**Node Creation:**

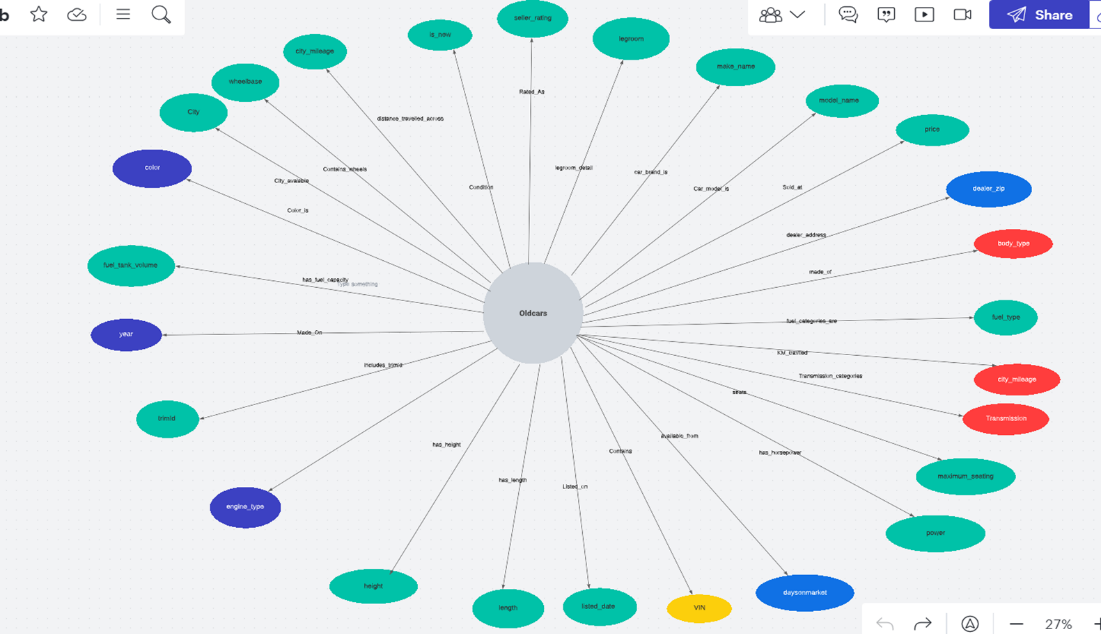


**Relationship Creation:**

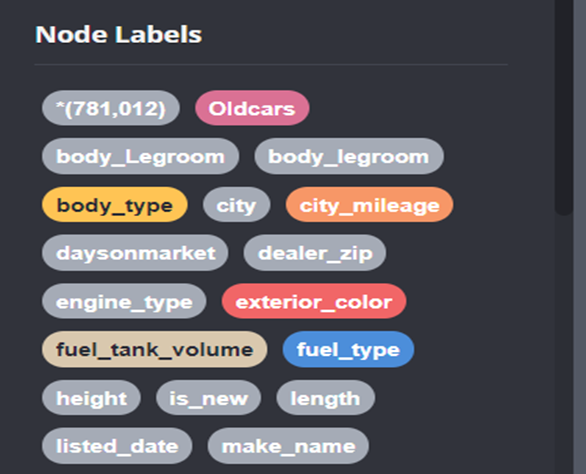


* Copy the cypher code(projectQuery\_script.txt) to create the constraints, nodes and relationships between them
* Run the code to create the graph database
* The scripts will run one by one and will take a few minutes due to the large dataset size The graph database should be prepared and can be tested by queries

**Graph Database:**



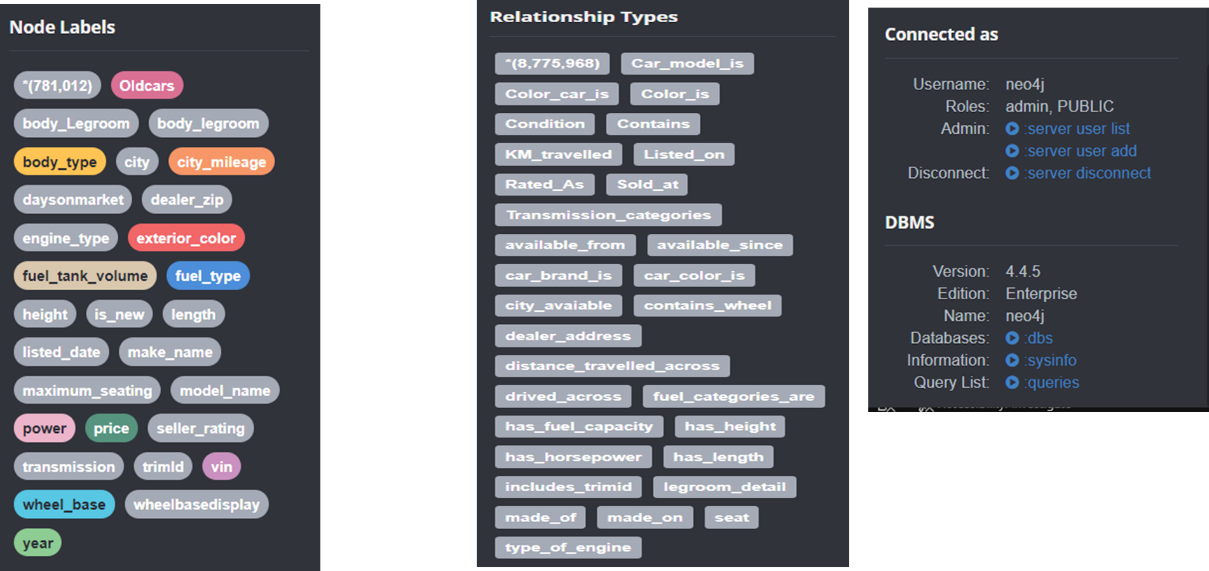
**Neo4j nodes:**







**Complete Nodes and relationship in Neo4j**



**Analytics Dashboard Documentation:**

1) Download Python if not already downloaded (follow instructions on https://www.python.org/downloads/ based on operating system

2) Install Anaconda Distribution (follow instructions on https://docs.anaconda.com/anaconda/install/index.html based on operating system)

3) Once installed, open Anaconda command prompt (anaconda3)

4) Install jupyter notebook by running command pip install notebook

5) Then, install dash framework by running pip command pip install jupyter-dash

6) After installing dash framework, install other required packages: · Pandas: pip install pandas · Matplotlib: pip install matplotlib · Plotly: pip install plotly=5.7.0 · NumPy: pip install numpy · Py2Neo: pip install py2neo

7) Open jupyter notebook by running command jupyter notebook on Anaconda command prompt

8) Jupyter Notebook web page should appear

9) Navigate to the folder containing dataset and app.ipynb. Ensure both files are in the same folder 10) Open app.ipynb on Jupyter Notebook

11)Run app.ipynb script

12)Wait until code fully loads and IP address appears at the bottom of the page

13) Navigate to IP address and wait for dashboard to fully load

14) Dashboard is displayed!

**Technical Metadata:**

Technical metadata consists of metadata that is associated with data transformation rules, data storage structures, semantic layers, and interface layers. Provides information on the format and structure of the data as needed by computer systems:





**Business Metadata:**

Business metadata is data that adds business context to other data. It provides information authored by business people and/or used by businesspeople.

Vin along with dealer zip, city and old Carid helps in identifying and filtering the used cars based on their location.

1.Number of used cars in the city

2.Various dealers in the city

3.Number of used cars sold and manufactured based on the dealer zip code.

·Length, Height, and Color are the dimensions of the used cars, and the car can be filtered based on this category by the business.

1.Filtering based on color, length, and height.

·Model name, make name, price, power, mileage, and year can be used by business to identify the total sales by model name and make name. The business can use the data to filter the number of used cars sold by their price, power, mileage, and year. Find the cars which have sold more that year on a certain price range.

·Body type, Engine type and legroom can be used by the business to find how many cars have which engine type and body type.

·Listing date, and Days on Market will be used by the business to find listing id and listing data of the vehicle on the website. Days on Market will be the days since the vehicle was first listed on the website.

·Fuel type and Fuel tank volume data can be used by businesses to find the car's fuel tanking capacity in gallons and filter the dominant fuel ingested by the vehicle.

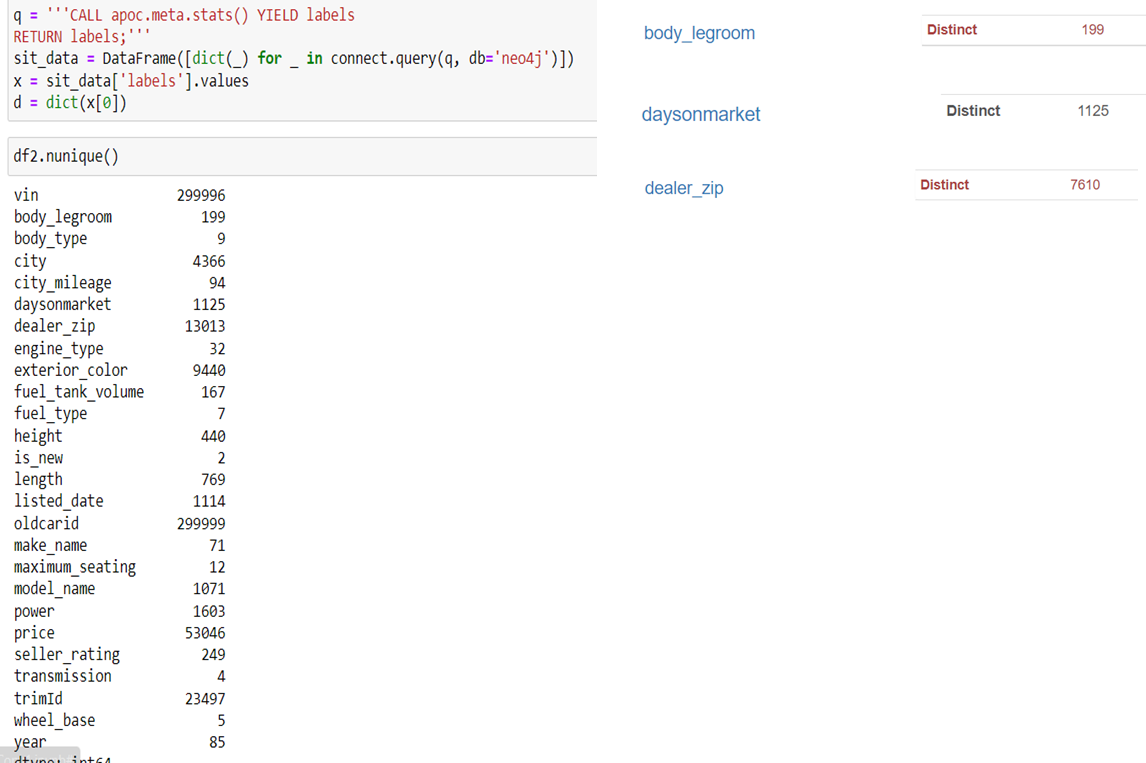
**Testing and Validation:**

**System Integration and User Acceptance Testing**

**Test 1: Check distinct values for each node**

* Distinct values for each node were calculated after creating the nodes and compared it to the values in the cleaned dataset file which was inserted into Neo4j.Made sure that the column counts for the unique value in the CSV file and the one loaded in the Neo4j database is the same.

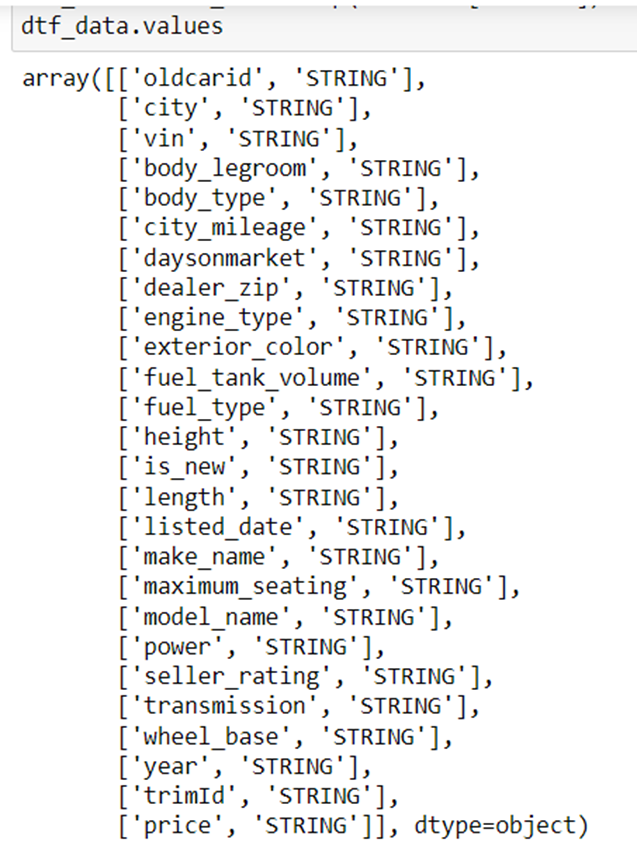
**Neo4j Count** **csv count**



**Test 2 : Verification of datatypes**

Verified data type of each node property by comparing data type of each node with cleaned dataset file.

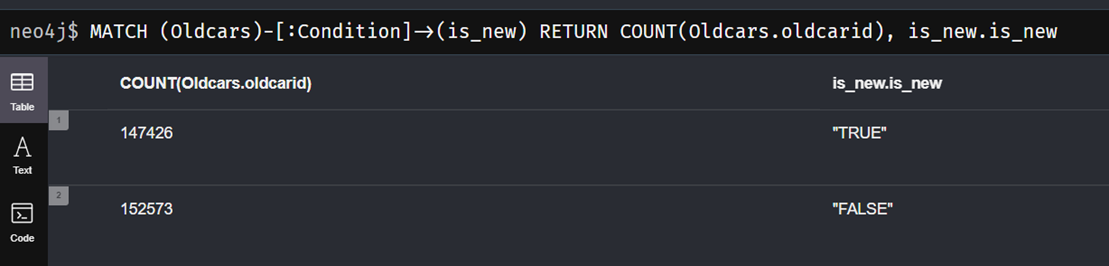
**Neo4j column data types**



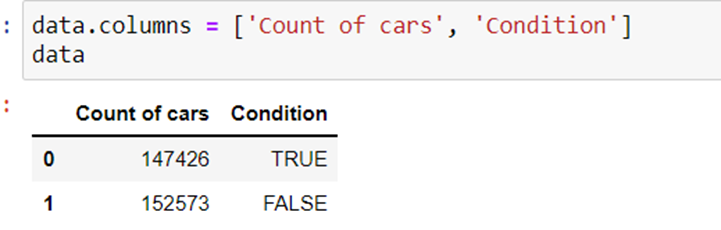
**Test 3 : Verifying counts in Neo 4j and the counts via python**

The query helps to get the count of the records in neo4j

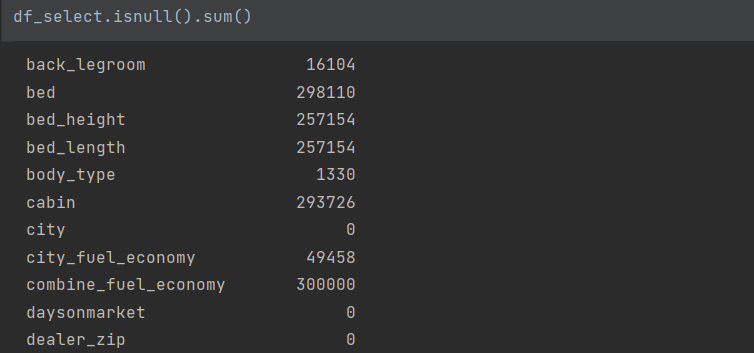
**Neo4j count**



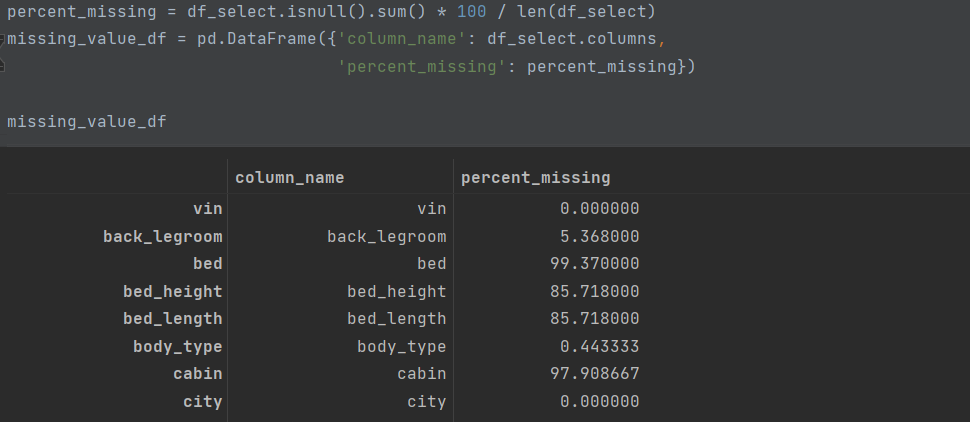
**Python Count**



**Test 4 : Check number of null values**



**Test 5 : Finding the missing value percentage in the dataset**



**Test 6 : Null values replaced**

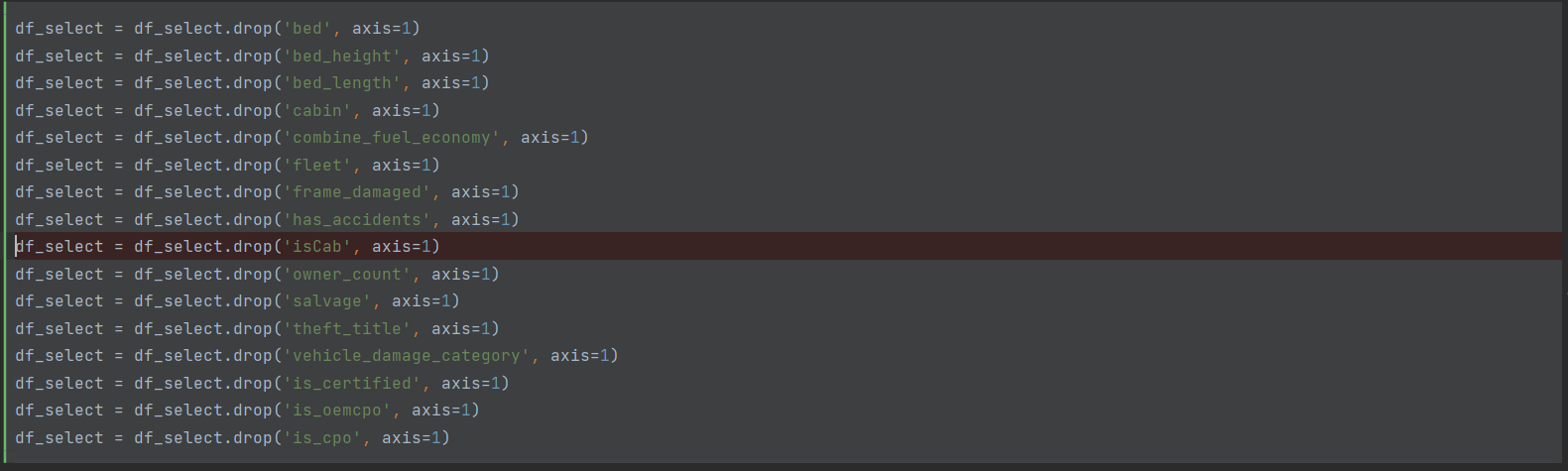
**Explanation**: These metrics were used in our data analysis and we wanted to account for

null values instead of just removing them from the dataset.



**Test 7: Dropped Columns Not Being Used for Analysis**

Explanation: Columns dropped were sigma values or scientific values that we could not fully understand scientifically for use in our data analysis.



**Test 8: Count of percent missing values will be zero**



**Neo4j Testing screenshots using Cypher Queries**

Steps for testing and verification of data using Neo4j:

● Copy the cypher code(neo4j\_testing\_script.txt) containing the test queries

● Run the code to view graphs and tables verifying data load to Neo4j

● The scripts will run one by one and will take a few minutes displaying the below results

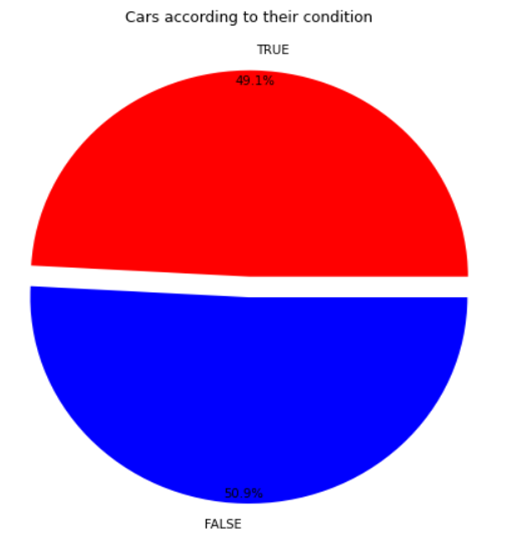
**Nodes Verification :**

* Distinct values for each nodes were calculated after creating the nodes and compared it to the values in the cleaned dataset file which was inserted into Neo4j
* Made sure that the column counts for the unique value in the CSV file and the one loaded in the Neo4j database is the same

**Dashboard Interpretation/Findings:**

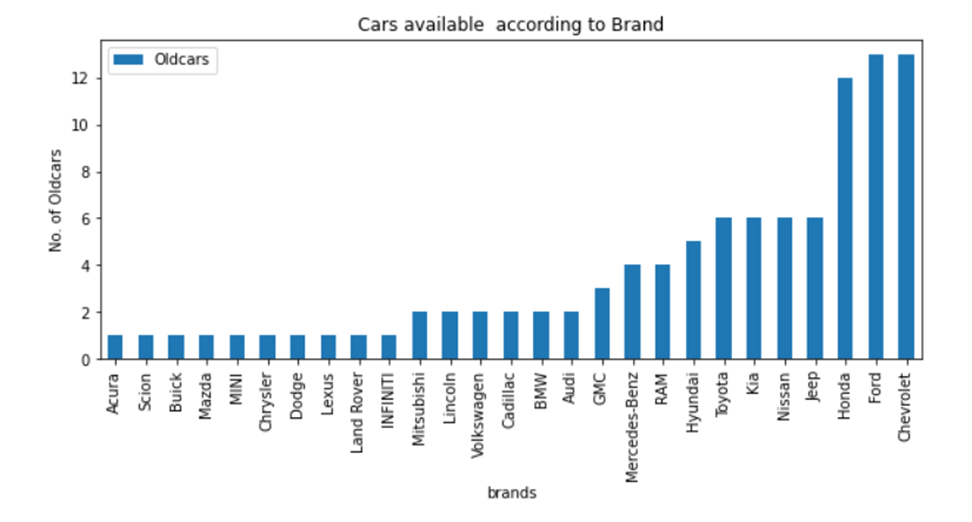
**1. Number of Cars according to the condition of the car**

* Highlights percentage of Used car condition that are in new condition
* The chart shows around 50.9% of cars are not new.
* This chart will give an idea to the buyer to get details regarding the cars that are in new condition at your price
* 49.1% of cars are in good condition, which means they are like new cars



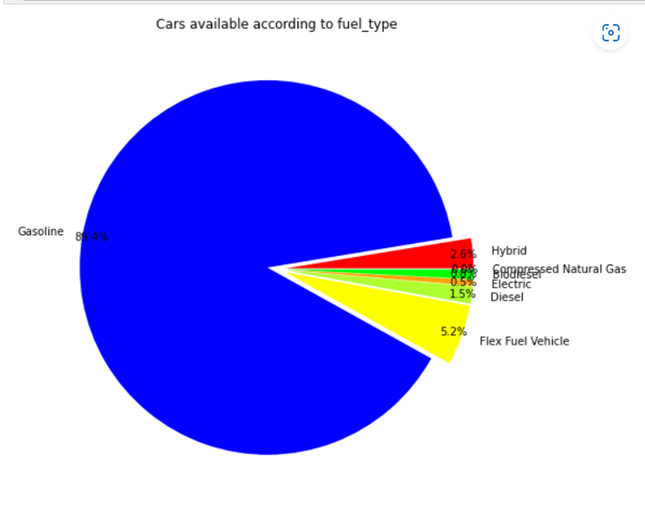
**2. Count of cars available according to year of manufacture**

* Dashboard depicts the count of used cars available in market for sale and the year of manufacture of car.
* This will give an idea regarding the cars available according to the year of manufacturing to the buyer so that he can decide of more to which kind of car he can go for.
* Using limit 100, we can find what is the count of the cars available which are manufactured recently and get more details regarding those cars.



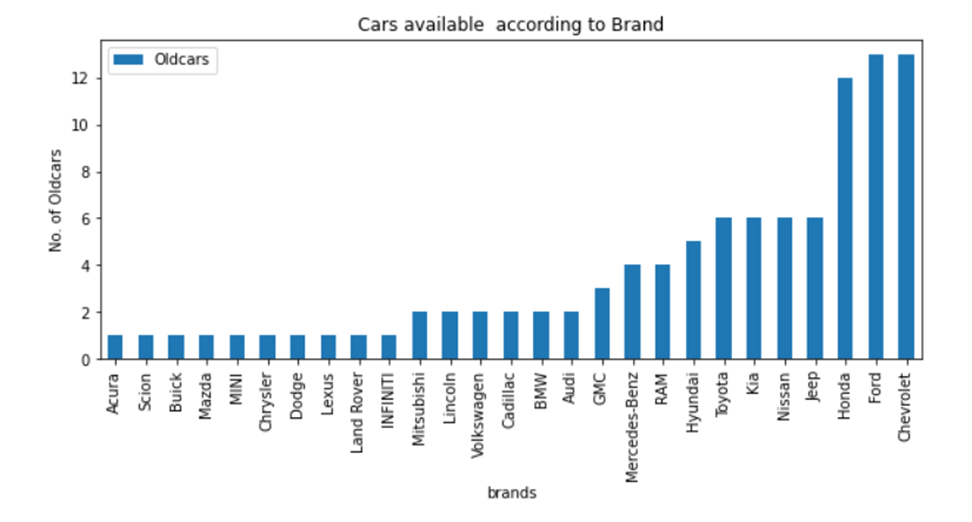
**3. Count of cars available according to fuel type**

* Highlights percentage of cars available according fuel type
* Chart shows that most 89.4%of cars are considered using gasoline and flex-fuel becomes 2nd highest as 5%
* This will give idea to buyers if there are any cars available according to their fuel preference.



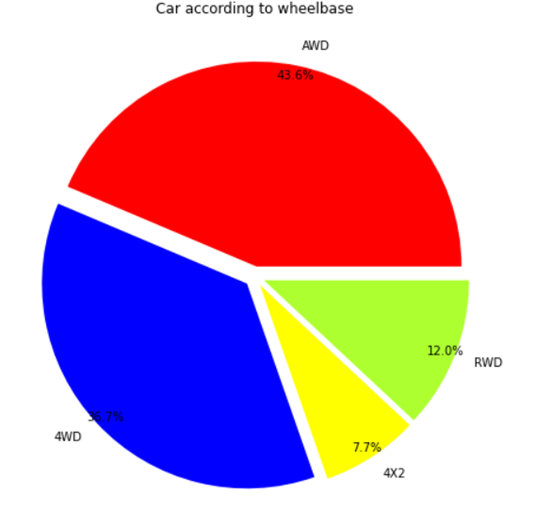
**4. Count of cars available according to Brands**

* Dashboard depicts the count of used cars available in market for sale and their brands
* This will give an idea regarding the cars available according to manufacturing brand so that they can decide as buyers have brand preference so it will help them to figure of the cars according to the brands
* Using limit 100, we can find what is the count of the cars available according to their brands



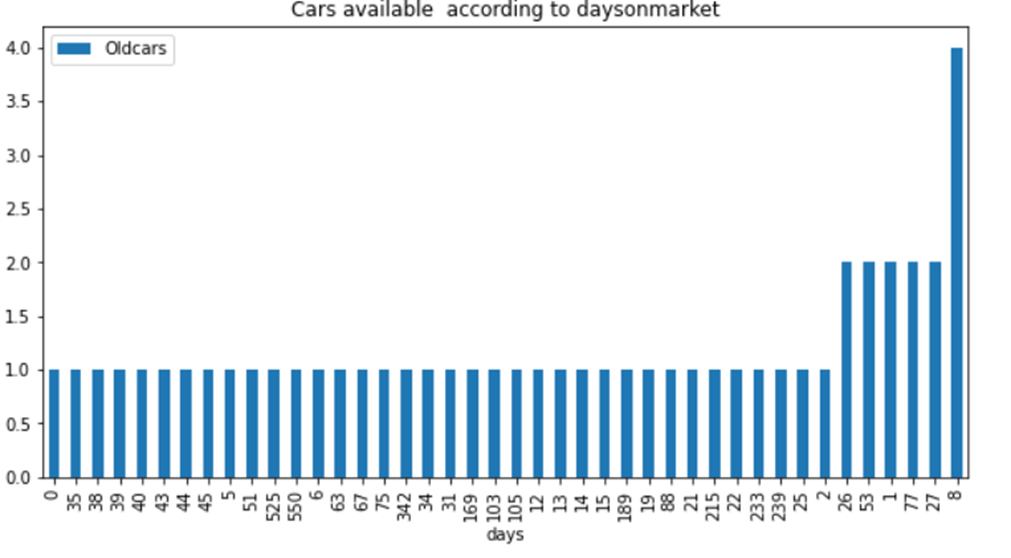
**5. Percentage of cars available according to Wheelbase**

* Highlights percentage of asteroids that are/are not potentially hazardous
* Chart shows that at most 43.6%of cars are considered having AWD wheels and 4WD being 36.7%
* This will give idea to buyers if there are any cars available according to their Wheelbase preference



**6. Count of cars available according to Days on market**

* Dashboard depicts the count of used cars available in market and the days they are available since markets.
* This helps the buyers to get idea about what are the latest cars available for purchasing also the cars which are market since long.
* For the cars which are in market since long even the seller can get an idea if his car will be in demand or not or else which location has that demand.
* Cars and the days on the markets are the main interpretation



**7. Count of cars available according to Seller rating**

* Highlights number of cars available with the seller having high ratings.
* If the seller is having good rating this helps the buyer and the seller both to get idea regarding the seller's service and they can go-ahead with that seller for either buying or selling cars
* Ratings help to get idea and decide as to which seller should proceed with the purchase
* Have below count of cars with sellers having with good and bad ratings



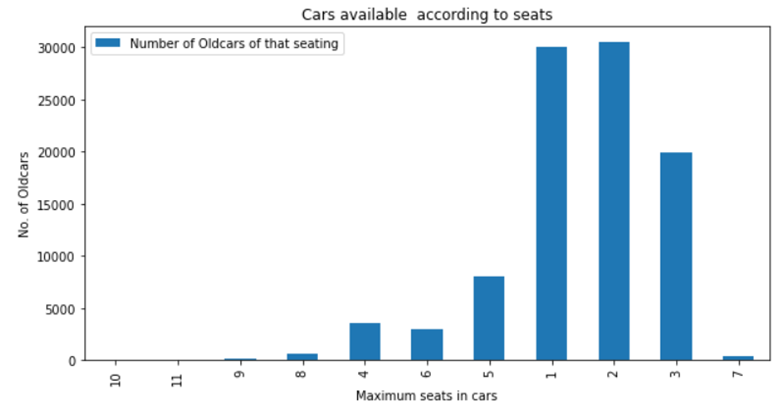
**8. Getting the car id and price of the car**

* Simply gives the idea of the price of the car
* Can sort the price desc and decide as which is the carid and then get details regarding the car may it be its color or type of interior or date of manufacturing
* Old car id and the price of the same for 10 sample old car is shown in the dashboard



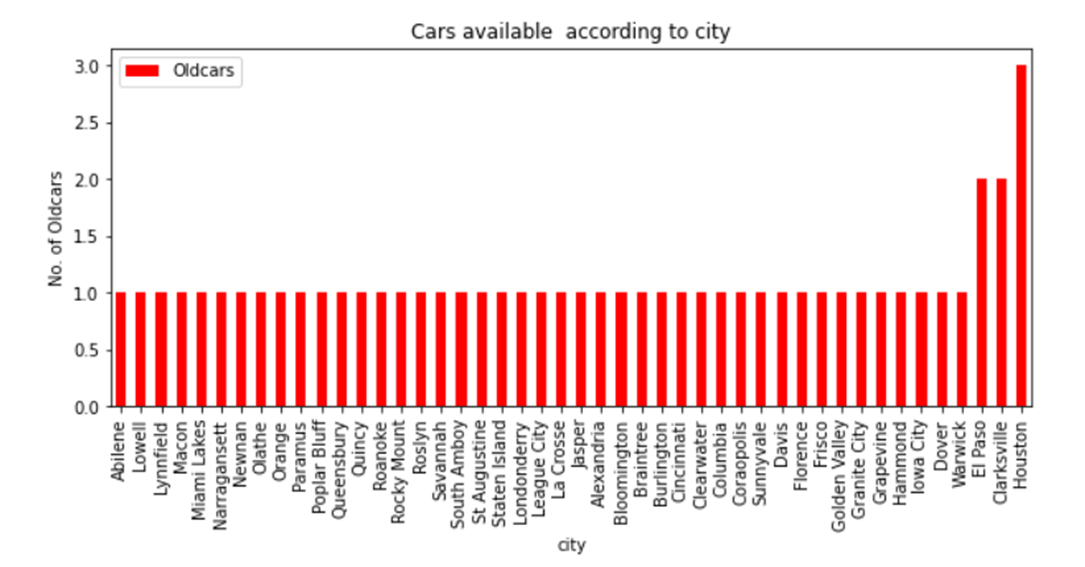
**9. Count of cars available according to seats in the car**

* Highlights number of cars available according to the seats in the cars
* As per the buyers requirements if the need 4/6 seaters cars he can get idea about the counts of cars
* Accordingly the buyer can decide if the cars which he wants to buy as per the number of seats is available in his area or not



**10. Count of cars available according to City**

* Dashboard depicts the count of used cars available in market according to your nearest city
* The dashboard have 50 cities and the counts of cars available around them.
* This gives idea to the buyer to get the counts of the car available in their area so that they can search for the cars available and its for them to buy the same.

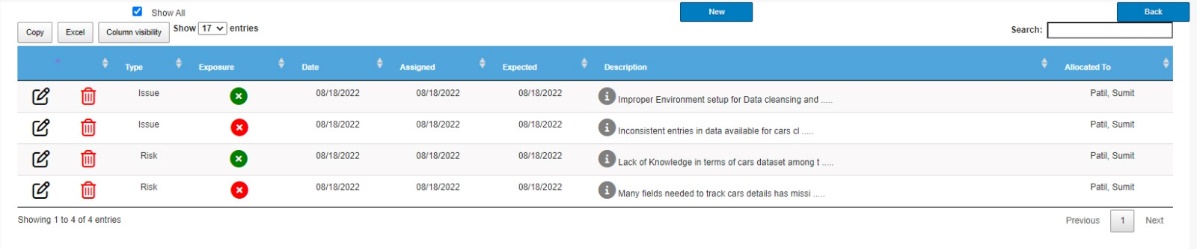


**Velero Screenshot:**

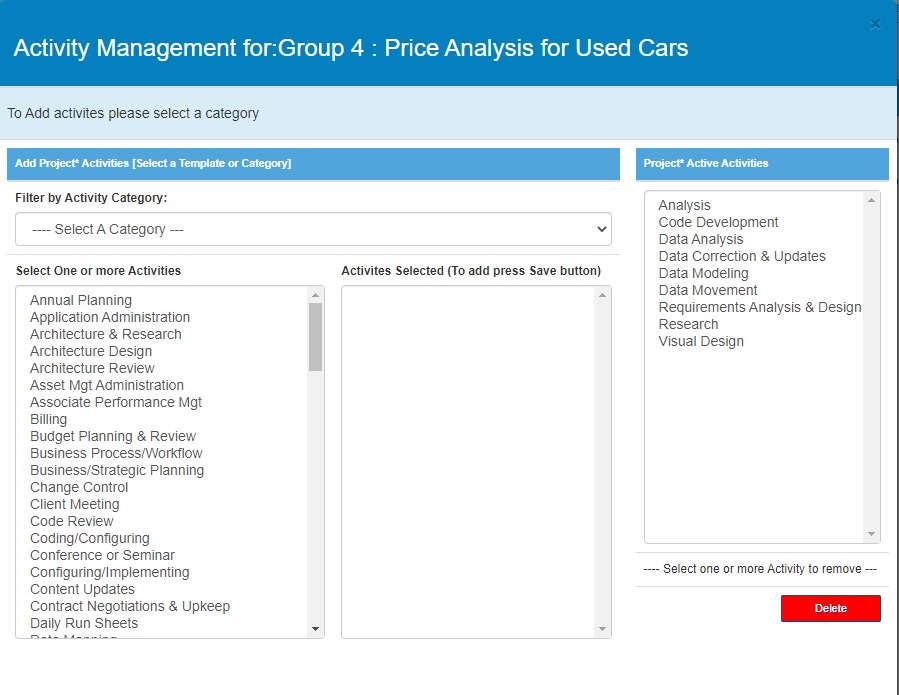
1. Resource Management



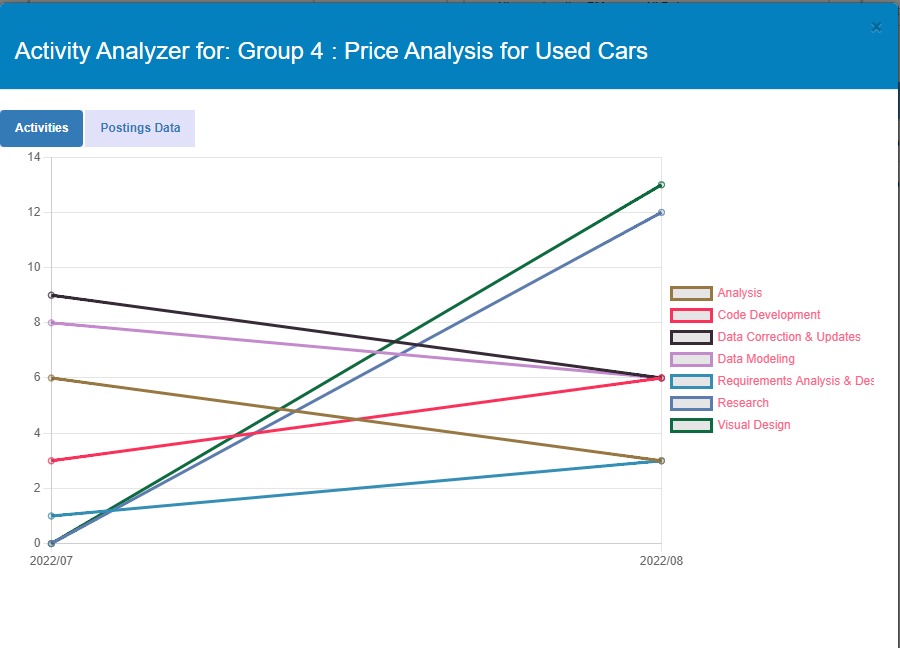
1. Risk & Issues:



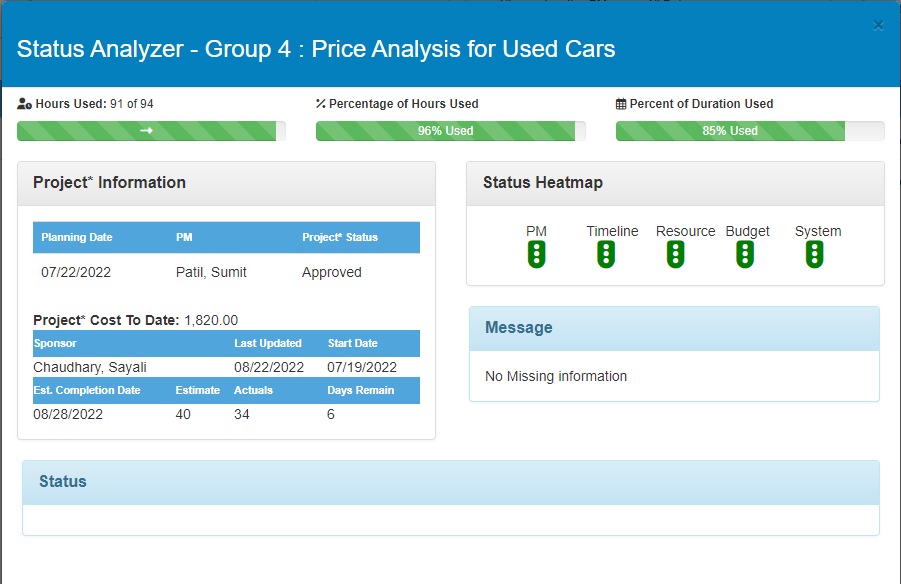
1. Activity Management



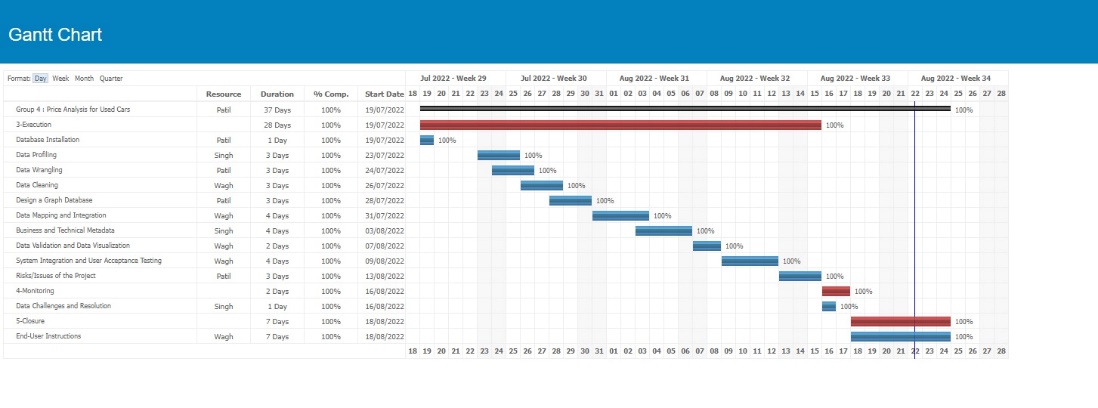
1. Activity Analyzer



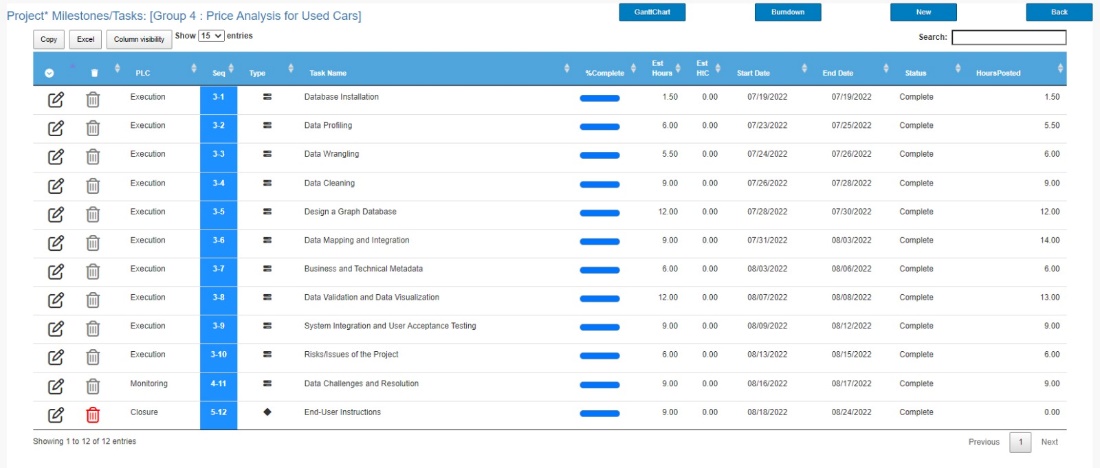
1. Status Analyzer



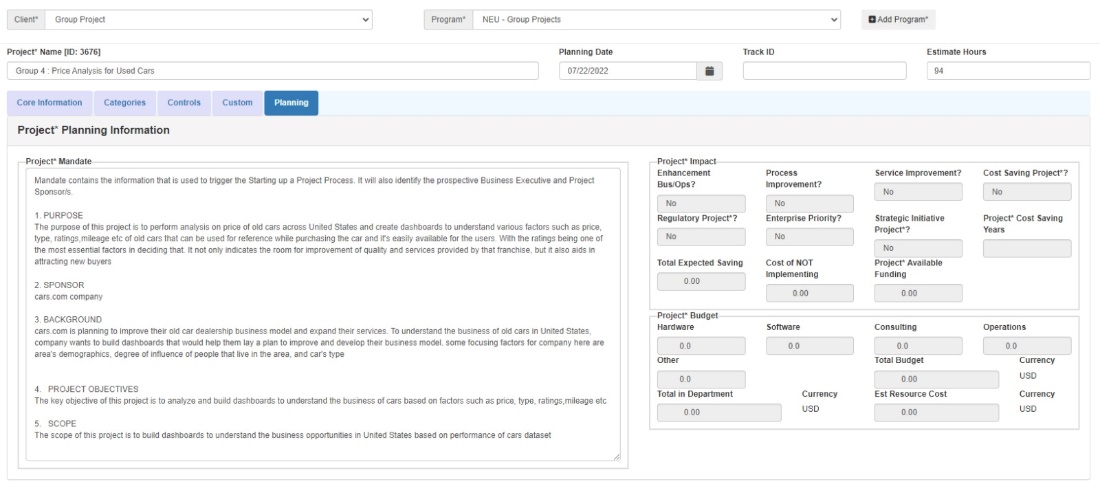
1. Gantt Chart



1. Management Plan



1. Mandates



1. Overview

