Grayson Nickel
9/1/2025

Cars, Prices and Deals

Project 1: Defining a Problem and Data Understanding

The problem

There are three questions that I want answered. First, what are the cheapest cars on the market? Second, what are the most expensive cars on the market? And finally, what cars give you the most power for the least amount of money? Inherently, this is more of a selfish question rather than something that will benefit everyone. I simply like good deals and want to see them. I'm sure there are other people asking these questions though.

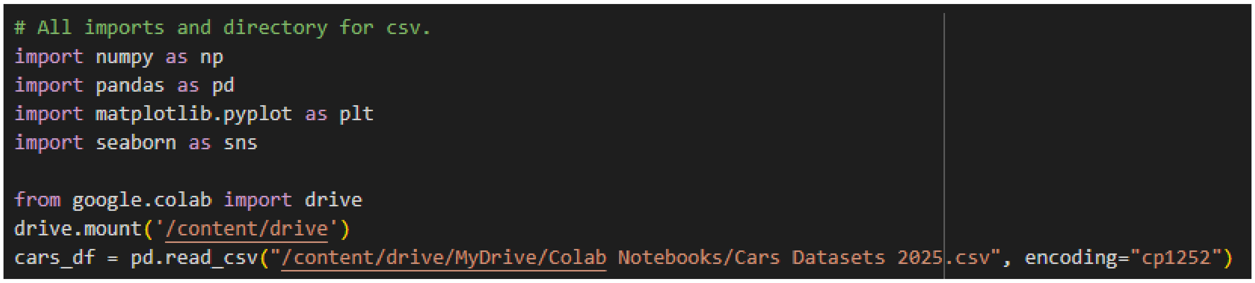
The Data

I found a data set for 2025 including over 1,200 cars. The dataset is called "Cars Datasets (2025)" posted by Abdul Malik on kaggle. This dataset is for free use as long as it's not malicious or for profit. This data set has 11 columns in total covering the following data:

Car Company Names: The manufacturer or brand of the car.
Car Models: The specific name or series of the car.
Engine Types: Information on engine specifications .
CC/Battery Capacity: Engine displacement in cubic centimeters or battery capacity for electric cars.
Horsepower (HP): The power output of the car's engine or motor.
Top Speed: The maximum speed the car can achieve.
0-100 km/h Performance: The time it takes for the car to accelerate from 0 to 100 km/h.
Price (in USD): The car's price listed in United States dollars.
Fuel Type: Specifies whether the car uses petrol, diesel, electricity, or hybrid fuel systems.
Seating Capacity: The number of passengers the car can accommodate.
Torque: The rotational force the engine generates.

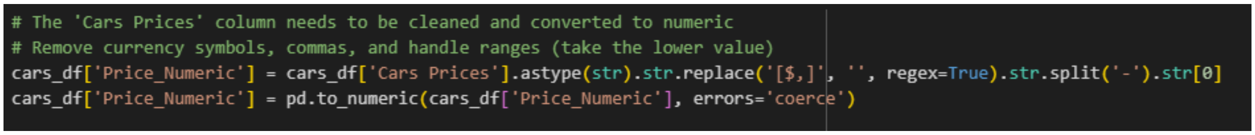
Pre-processing

My first step was to set up my environment with all the necessary imports and directories for the data set.



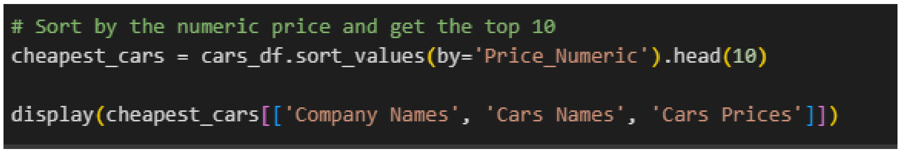
```
# All imports and directory for csv.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

from google.colab import drive
drive.mount('/content/drive')
cars_df = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/Cars Datasets 2025.csv", encoding="cp1252")
```

Next, I had to convert the values of “Cars Prices” into something more usable for calculations. All the dollar signs and commas were an issue. I assigned the new values to “Price_Numeric”.

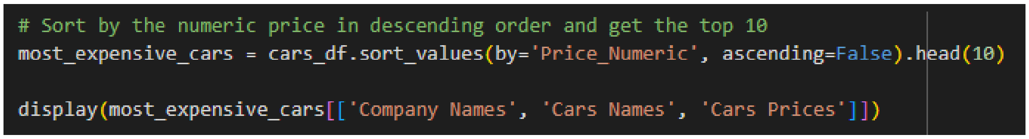


```
# The 'Cars Prices' column needs to be cleaned and converted to numeric
# Remove currency symbols, commas, and handle ranges (take the lower value)
cars_df['Price_Numeric'] = cars_df['Cars Prices'].astype(str).str.replace('[$,]', '', regex=True).str.split('-').str[0]
cars_df['Price_Numeric'] = pd.to_numeric(cars_df['Price_Numeric'], errors='coerce')
```

Next I displayed the 10 cheapest and 10 most expensive cars on the list in order.

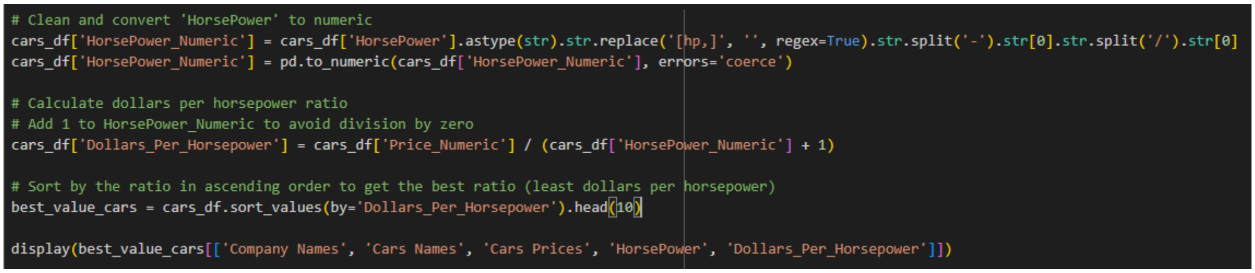


```
# Sort by the numeric price and get the top 10
cheapest_cars = cars_df.sort_values(by='Price_Numeric').head(10)

display(cheapest_cars[['Company Names', 'Cars Names', 'Cars Prices']])
```



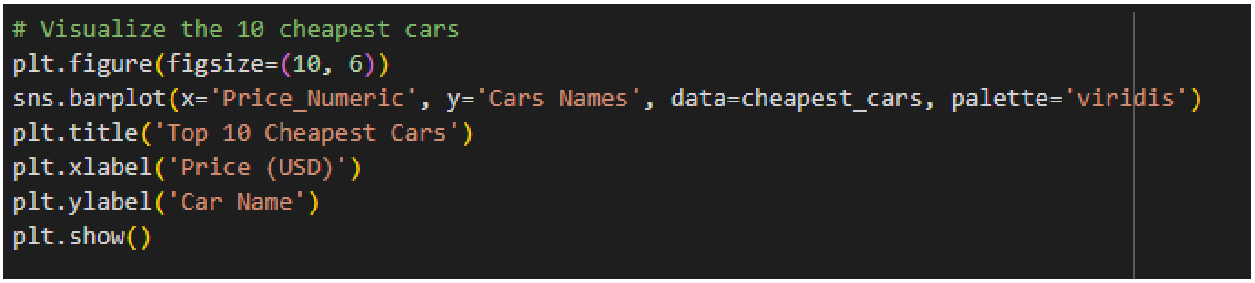
```
# Sort by the numeric price in descending order and get the top 10
most_expensive_cars = cars_df.sort_values(by='Price_Numeric', ascending=False).head(10)

display(most_expensive_cars[['Company Names', 'Cars Names', 'Cars Prices']])
```

Next, I “cleaned” the “Horsepower” the same way I cleaned the price. Then using the new “Price_Numeric” and “Horsepower_Numeric”, I was able to divide the price by the horsepower to create “Dollars_Per_Horsepower”. Using “Dollars_Per_Horsepower”, I was able to display the cars with the least cost per horsepower.

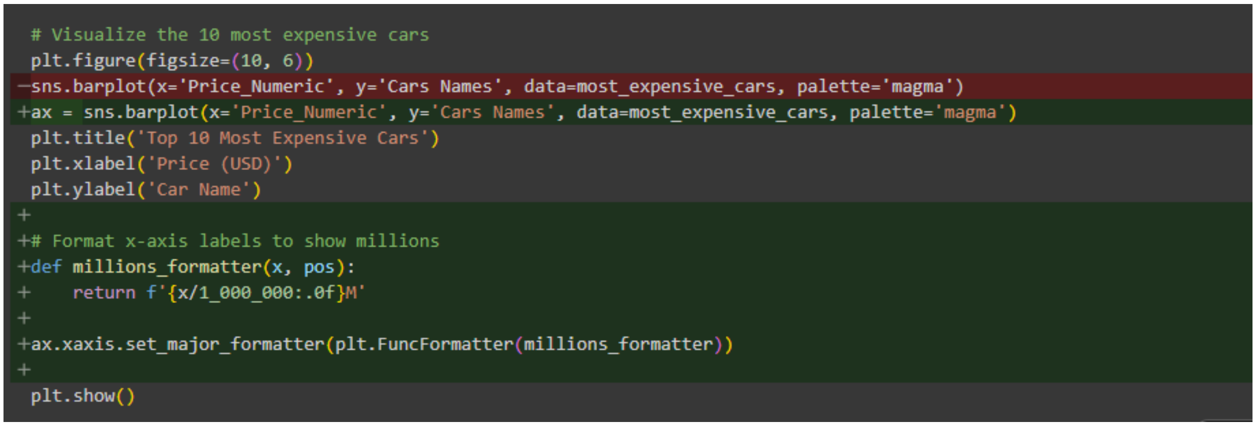


```
# Clean and convert 'HorsePower' to numeric
cars_df['HorsePower_Numeric'] = cars_df['HorsePower'].astype(str).str.replace('[hp,]', '', regex=True).str.split('-').str[0].str.split('/').str[0]
cars_df['HorsePower_Numeric'] = pd.to_numeric(cars_df['HorsePower_Numeric'], errors='coerce')

# Calculate dollars per horsepower ratio
# Add 1 to HorsePower_Numeric to avoid division by zero
cars_df['Dollars_Per_Horsepower'] = cars_df['Price_Numeric'] / (cars_df['HorsePower_Numeric'] + 1)

# Sort by the ratio in ascending order to get the best ratio (least dollars per horsepower)
best_value_cars = cars_df.sort_values(by='Dollars_Per_Horsepower').head(10)

display(best_value_cars[['Company Names', 'Cars Names', 'Cars Prices', 'HorsePower', 'Dollars_Per_Horsepower']])
```

With all the questions technically answered at this point, I went on to visualize the answers to all three of my questions.

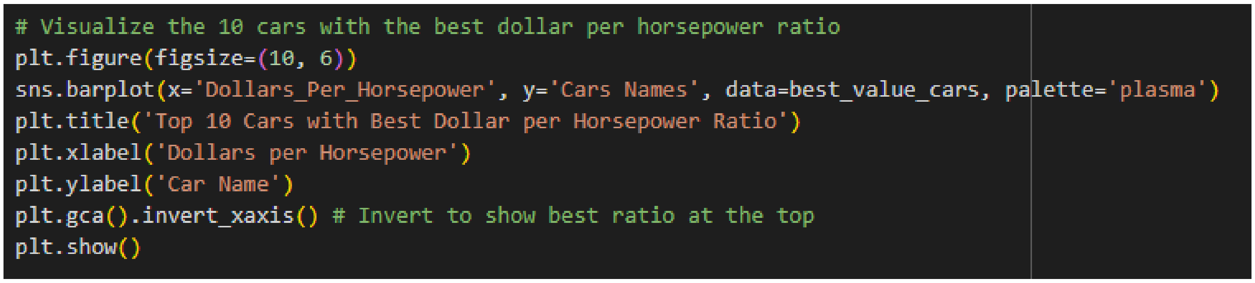


```
# Visualize the 10 cheapest cars
plt.figure(figsize=(10, 6))
sns.barplot(x='Price_Numeric', y='Cars Names', data=cheapest_cars, palette='viridis')
plt.title('Top 10 Cheapest Cars')
plt.xlabel('Price (USD)')
plt.ylabel('Car Name')
plt.show()
```



```
 # Visualize the 10 most expensive cars
 plt.figure(figsize=(10, 6))
-sns.barplot(x='Price_Numeric', y='Cars Names', data=most_expensive_cars, palette='magma')
+ax = sns.barplot(x='Price_Numeric', y='Cars Names', data=most_expensive_cars, palette='magma')
 plt.title('Top 10 Most Expensive Cars')
 plt.xlabel('Price (USD)')
 plt.ylabel('Car Name')
+
+# Format x-axis labels to show millions
+def millions_formatter(x, pos):
+    return f'{x/1_000_000:.0f}M'
+
+ax.xaxis.set_major_formatter(plt.FuncFormatter(millions_formatter))
+
 plt.show()
```

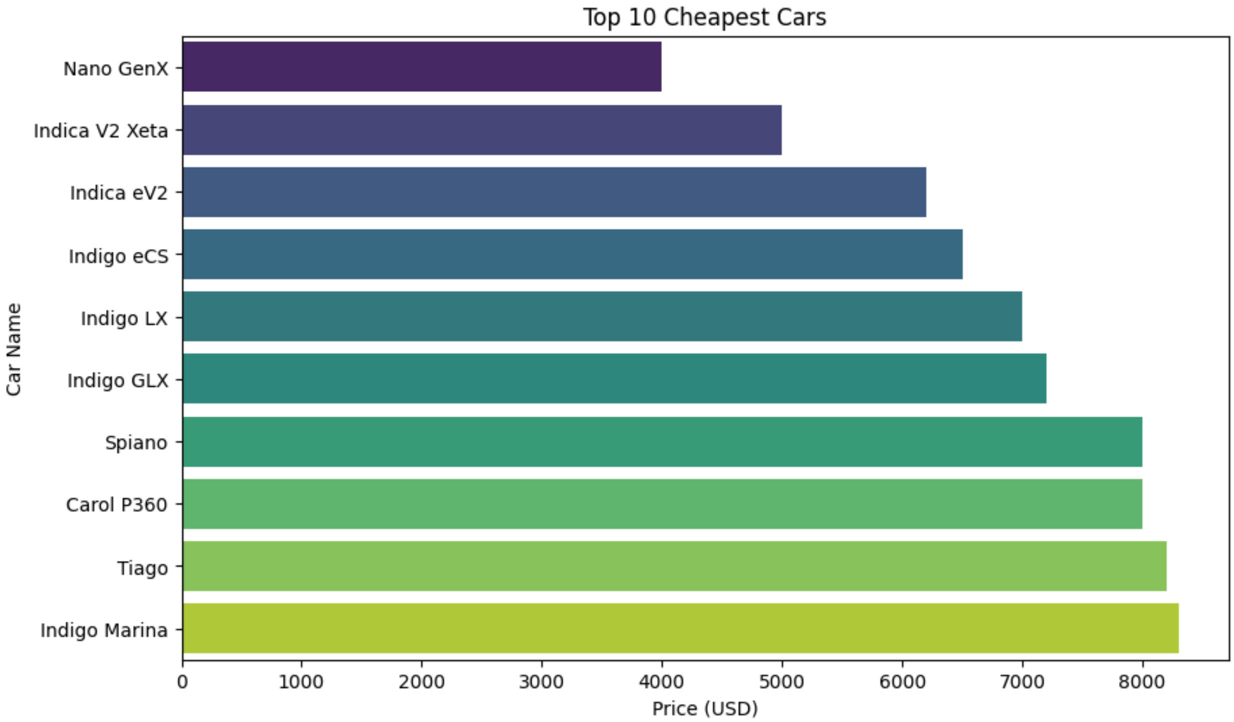
I received help with Gemini for the 10 most expensive cars graph as I was having issues getting it to show price in millions rather than a strange decimal value. You can see the recommendations it made highlighted in green.

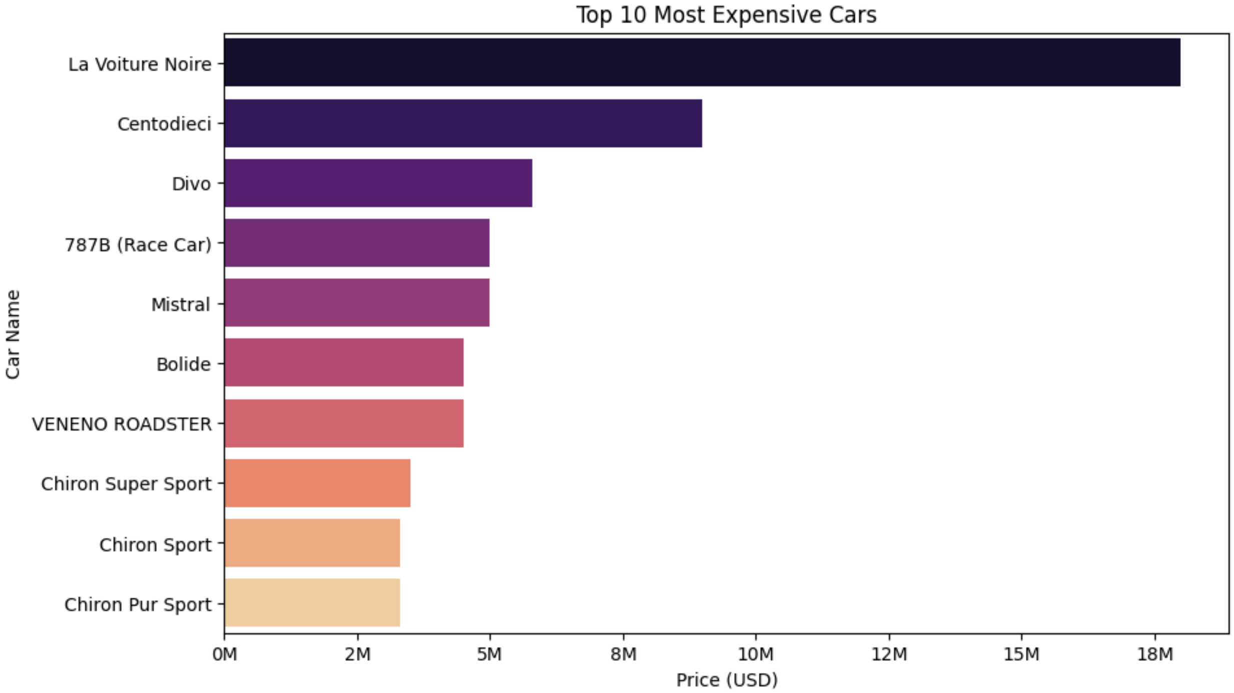


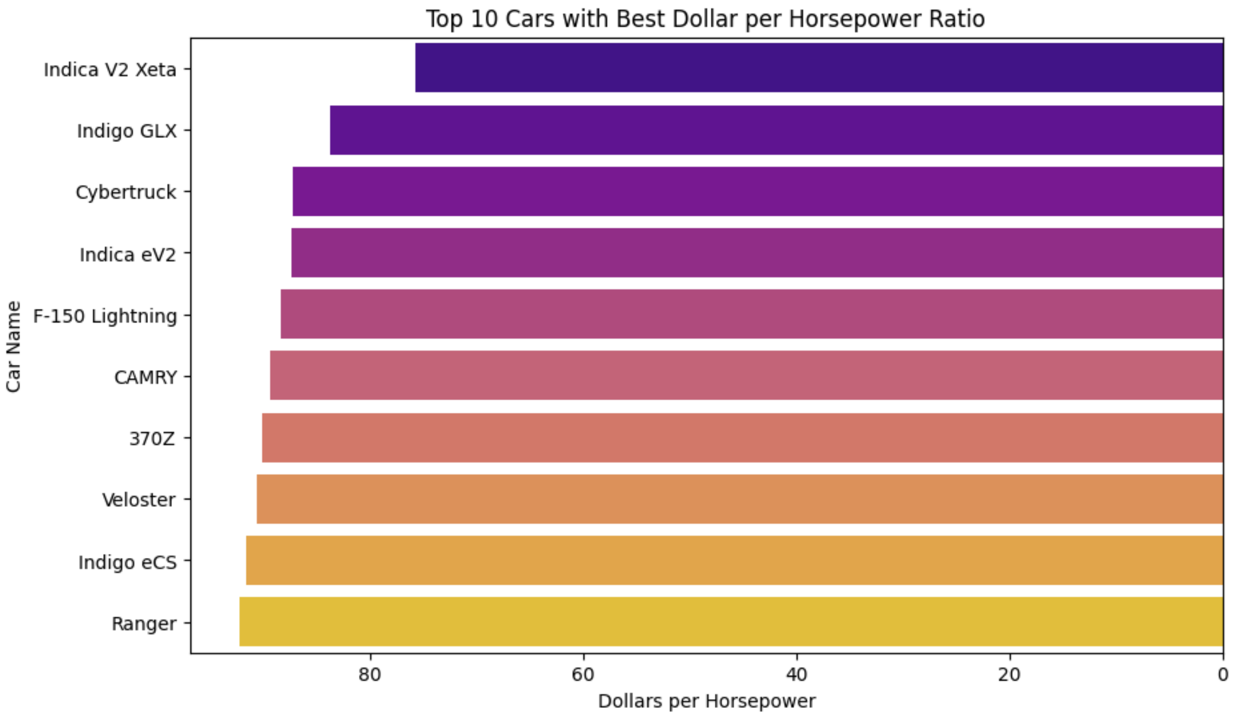
```
# Visualize the 10 cars with the best dollar per horsepower ratio
plt.figure(figsize=(10, 6))
sns.barplot(x='Dollars_Per_Horsepower', y='Cars Names', data=best_value_cars, palette='plasma')
plt.title('Top 10 Cars with Best Dollar per Horsepower Ratio')
plt.xlabel('Dollars per Horsepower')
plt.ylabel('Car Name')
plt.gca().invert_xaxis() # Invert to show best ratio at the top
plt.show()
```

Data Understanding/Visualization

While exploring the cheapest cars, I was surprised to see that I didn't recognize a single car on the cheapest cars list. I was expecting to have at least heard of a couple. Another thing that caught my attention is that I actually own one of the cars on the 10 best dollar per horsepower list. That car would be the toyota camry, however mine is not a 2025.


Top 10 Cheapest Cars

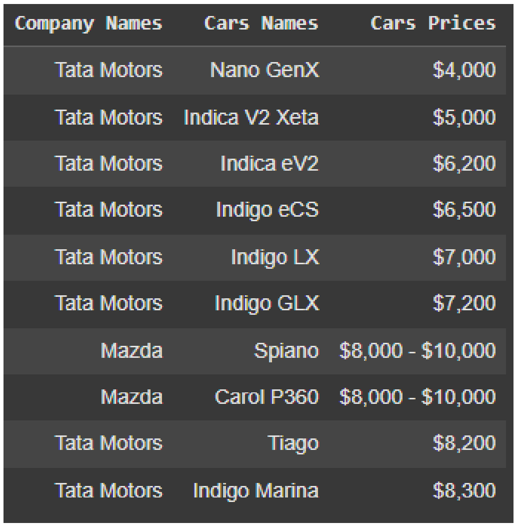

Top 10 Most Expensive Cars



Storytelling

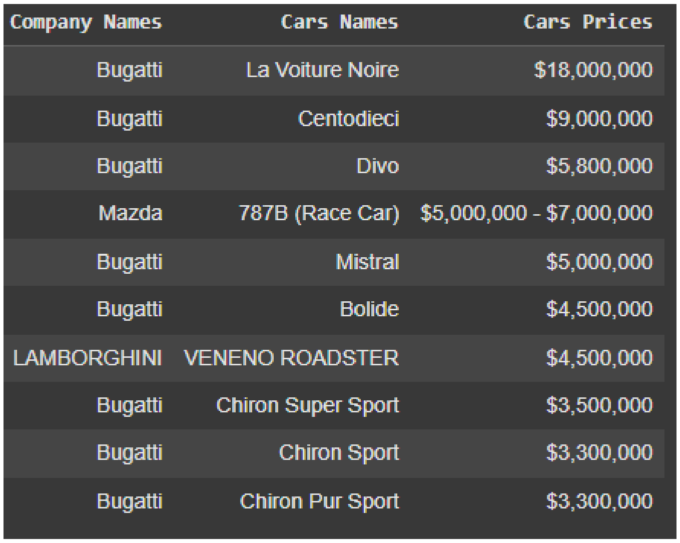
All of my initial questions have been answered upon completing the display() methods and visualizations. The visualizations are labeled, display all the necessary information about the cars and have a color gradient to help differentiate things.

Q1 - The 10 cheapest cars of 2025 are:



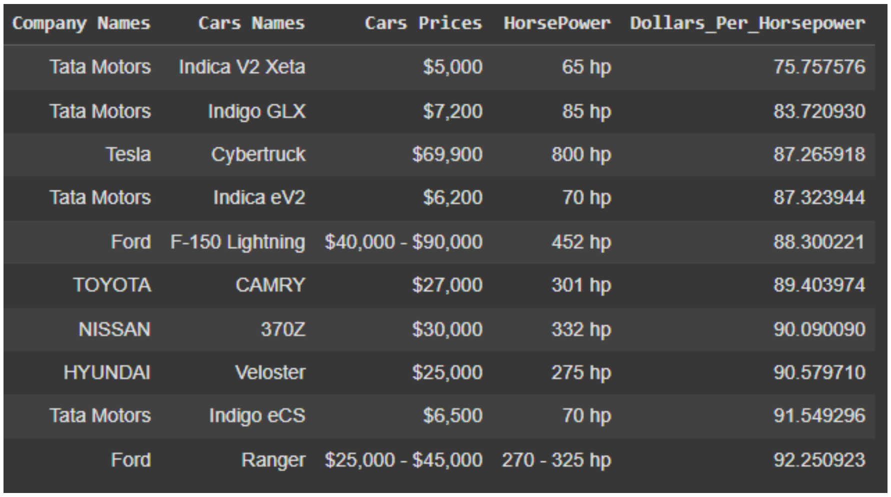
Company Names	Cars Names	Cars Prices
Tata Motors	Nano GenX	$4,000
Tata Motors	Indica V2 Xeta	$5,000
Tata Motors	Indica eV2	$6,200
Tata Motors	Indigo eCS	$6,500
Tata Motors	Indigo LX	$7,000
Tata Motors	Indigo GLX	$7,200
Mazda	Spiano	$8,000 - $10,000
Mazda	Carol P360	$8,000 - $10,000
Tata Motors	Tiago	$8,200
Tata Motors	Indigo Marina	$8,300

Q2 - The 10 most expensive cars of 2025 are:



Company Names	Cars Names	Cars Prices
Bugatti	La Voiture Noire	$18,000,000
Bugatti	Centodieci	$9,000,000
Bugatti	Divo	$5,800,000
Mazda	787B (Race Car)	$5,000,000 - $7,000,000
Bugatti	Mistral	$5,000,000
Bugatti	Bolide	$4,500,000
LAMBORGHINI	VENENO ROADSTER	$4,500,000
Bugatti	Chiron Super Sport	$3,500,000
Bugatti	Chiron Sport	$3,300,000
Bugatti	Chiron Pur Sport	$3,300,000

Q3 The highest horsepower yield per dollar of 2025 are:



Company Names	Cars Names	Cars Prices	HorsePower	Dollars_Per_Horsepower
Tata Motors	Indica V2 Xeta	$5,000	65 hp	75.757576
Tata Motors	Indigo GLX	$7,200	85 hp	83.720930
Tesla	Cybertruck	$69,900	800 hp	87.265918
Tata Motors	Indica eV2	$6,200	70 hp	87.323944
Ford	F-150 Lightning	$40,000 - $90,000	452 hp	88.300221
TOYOTA	CAMRY	$27,000	301 hp	89.403974
NISSAN	370Z	$30,000	332 hp	90.090090
HYUNDAI	Veloster	$25,000	275 hp	90.579710
Tata Motors	Indigo eCS	$6,500	70 hp	91.549296
Ford	Ranger	$25,000 - $45,000	270 - 325 hp	92.250923

I learned that there are many other kinds of cars out there than just what's available in the US. Especially when it comes to the cheaper market when certain federal safety and emissions standards don't need to be met. A prime example of this would be the Tata Motors Indica V2 Xeta which is the cheapest car from 2025 on this list. If you could get something brand new like that here in the US for only $5,000, that would be incredible.

Impact Section

My answered questions could benefit people looking to find the highest powered car for the money or those who are just curious about what's out there. A negative could be that some of the cars on this list might be rare or not available in the US. That paired with other vehicles having missing data could hurt the results. Now, one harmful thing I thought about from my findings is something I mentioned earlier about the Tata Motors Indica V2 Xeta. If you are so focused on price and decide or buy one of the cheap vehicles on this list without any extra thought (wherever you may live), you could be harmed in the long run. Specifically, I am referring to the lack of federal safety standards for some of these cheaper cars not intended for the US market. I really doubt the Tata Motors Indica V2 Xeta is capable of holding up nearly as well in a high speed collision versus something you may find in a dealership here in the states. With that in mind, one thing I think this dataset lacks is some sort of data on the cars safety like crash test results or something along those lines.

References

The dataset: https://www.kaggle.com/datasets/abdulmalik1518/cars-datasets-2025
Seaborn API: https://seaborn.pydata.org/api.html
Matplotlib Library: https://matplotlib.org/
Copilot for error handling (Specifically for 10 most expensive cars visual)

The Code

All of the project files including the code and csv can be found here:
https://github.com/GMNICKEL/Projects/tree/main/Python%20datamining%20project%201