DATA & AI SPECIALIST MID-TERM EXAMINATION

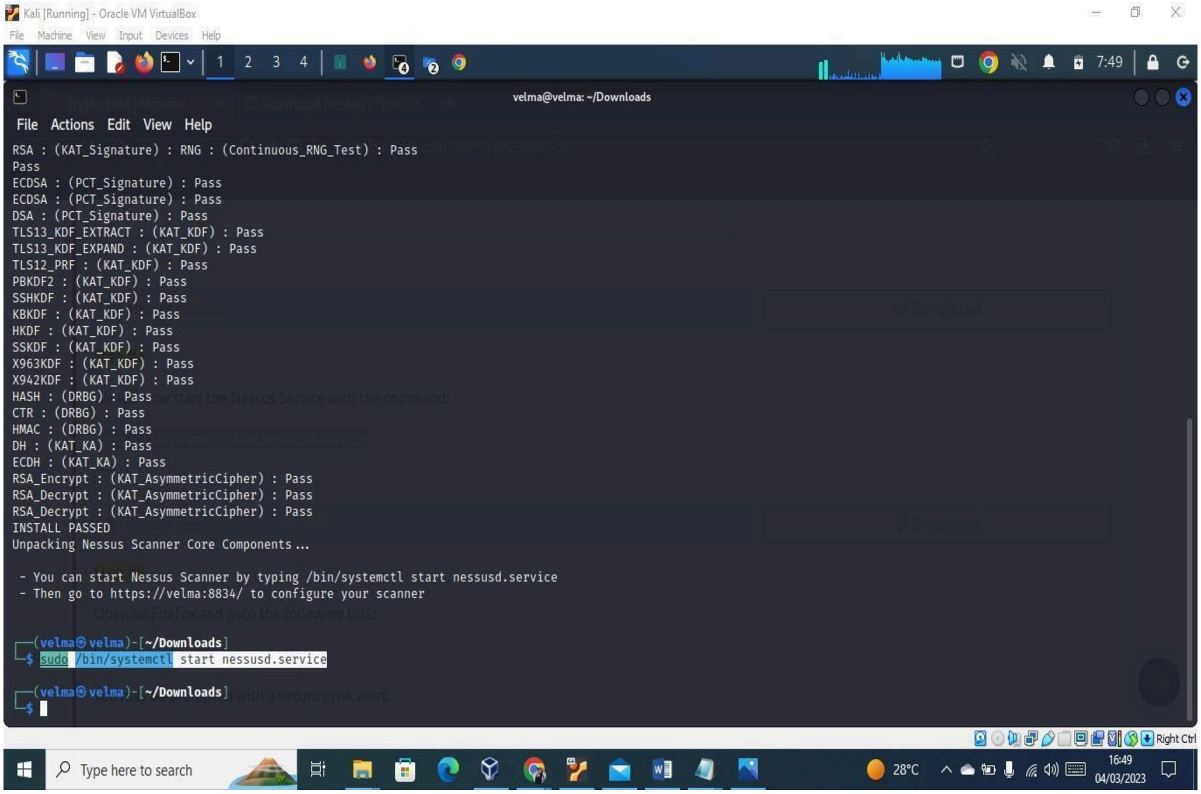
MID EXAMINATION

TIME ALLOWED: 2 HOURS

TOTAL: 100 MARKS

Instructions:

1. Answer ALL questions
2. The exam should NOT be worked on in groups or with assistance from others.
3. Use this file as your write-up reporting template as you complete each task outlined and answer the questions.
4. Rename this file with your full names and Cyber Shujaa ID.
5. Once you have completed your work, save the file and upload it for marking.
6. Before leaving the exam, ensure you have uploaded the correct file capturing all the work you have submitted for marking.
7. Ensure you compile a detailed report write-up that outlines your approach to addressing the various exam challenges. Ensure that your write up is authentic. Show screenshots of the working for all answers showing how you got your answers.
8. The screen shots should capture your full screen and display the command you ran to get the answer. Include a taskbar showing your machine taskbar and time stamp
9. MCQS will be available from 6:30 to 7pm.


Kali [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
velma@velma: ~/Downloads
File Actions Edit View Help
7:49
Type here to search
28°C
16:49
04/03/2023

```
RSA : (KAT_Signature) : RNG : (Continuous_RNG_Test) : Pass
Pass
ECDSA : (PCT_Signature) : Pass
ECDSA : (PCT_Signature) : Pass
DSA : (PCT_Signature) : Pass
TLS13_KDF_EXTRACT : (KAT_KDF) : Pass
TLS13_KDF_EXPAND : (KAT_KDF) : Pass
TLS12_PRF : (KAT_KDF) : Pass
PBKDF2 : (KAT_KDF) : Pass
SSHKDF : (KAT_KDF) : Pass
KBKDF : (KAT_KDF) : Pass
HKDF : (KAT_KDF) : Pass
SSKDF : (KAT_KDF) : Pass
X963KDF : (KAT_KDF) : Pass
X942KDF : (KAT_KDF) : Pass
HASH : (DRBG) : Pass
CTR : (DRBG) : Pass
HMAC : (DRBG) : Pass
DH : (KAT_KA) : Pass
ECDH : (KAT_KA) : Pass
RSA_Encrypt : (KAT_AsymmetricCipher) : Pass
RSA_Decrypt : (KAT_AsymmetricCipher) : Pass
RSA_Decrypt : (KAT_AsymmetricCipher) : Pass
INSTALL PASSED
Unpacking Nessus Scanner Core Components...

 - You can start Nessus Scanner by typing /bin/systemctl start nessusd.service
 - Then go to https://velma:8834/ to configure your scanner

┌──(velma㉿velma)-[~/Downloads]
└─$ sudo /bin/systemctl start nessusd.service

┌──(velma㉿velma)-[~/Downloads]
└─$ 
```

Dataset:

- Download the required dataset from this link:Used Cars Dataset. You will use the `used_cars.csv` file.

Instructions:

- Answer **ALL** questions.
- This exam must **NOT** be worked on in groups or with assistance from others.
- Compile a detailed report in a PDF document that outlines your approach, code, and findings. **Screenshots are mandatory** for all tasks to show your process. Screenshots must be clear and show the full screen, including the taskbar with the time.

SECTION B: PRACTICAL EXAMINATION

Scenario: Analyzing the Used Car Market

You are a Junior Data Analyst at **"AutoAnalytics Inc.,"** a consulting firm that advises used car marketplaces. Your firm has been hired to analyze the factors that influence the price of used cars based on data from `cars.com`. Your task is to clean the provided dataset, identify key trends, and build an interactive tool that can help dealership managers understand market pricing.

The insights from your analysis will help dealerships optimize their inventory, set competitive prices, and advise customers effectively. The primary column of interest is `price`.

TASKS

Part 1: Data Parsing (10 Marks)

Your team receives data from various sources. One source provides data as a single text string.

Data Snippet: `"brand: Ford, model: F-150, model_year: 2023, milage: 2,823 mi., price: 45000; brand: BMW, model: 740 iL, model_year: 2001, milage: 242,000 mi., price: 8500; brand: Tesla, model: Model X, model_year: 2020, milage: 34,000 mi., price: 75000"`

1. Write a Python script to parse this string.
2. Extract the details for each car (brand, model, year, mileage, price).
3. Load the extracted data into a clean pandas DataFrame.

Part 2: Data Wrangling & Preparation (30 Marks)

For this section, use the `used_cars.csv` file you downloaded. Note that the mileage column is misspelled as `milage`.

1. **(5 Marks)** Load the dataset using pandas and display its first 10 rows and a summary of its structure (`.info()`).
2. **(5 Marks)** The `milage` column is a string (e.g., "51,000 mi."). Clean this column by removing commas and " mi.", then convert it to a numerical data type (integer).
3. **(5 Marks)** The `accident` column is categorical text. Convert it into a **boolean** column (True/False) or a **binary** column (1/0) where a value of `True` or `1` indicates that an accident was reported.
4. **(5 Marks)** Create a new column called **`car_age`** by subtracting the `model_year` from the current year (e.g., 2025).
5. **(10 Marks)** The `engine` column is complex (e.g., "300.0HP 3.7L V6..."). Create a new column called **`horsepower`** by extracting only the horsepower number (the number that appears before "HP").
    - Use regular expressions (`regex`) for this task.
    - Ensure the new `horsepower` column is a numerical type.
    - If horsepower is not found for a row, the value should be `NaN` (Not a Number). Handle these resulting missing values by filling them with the median horsepower.

Part 3: Exploratory Data Analysis (EDA) (35 Marks)

Use Python's `matplotlib` and `seaborn` libraries for this section. The target variable for analysis is `price`.

1. **(10 Marks) Univariate Analysis:**
    - Create a histogram to show the distribution of car `price`.
    - Create a boxplot to identify outliers in the `milage` column.
2. **(15 Marks) Bivariate Analysis:**
    - Create a scatter plot to visualize the relationship between `milage` (x-axis) and `price` (y-axis).

- Create a box plot to compare the distribution of `price` for cars with an accident history versus those with none.
- Create a violin plot showing the relationship between `fuel_type` and `price`.

3. **(10 Marks)** Summarize your findings by listing and explaining **at least three meaningful insights** from your EDA. (e.g., "The price of a car decreases as mileage increases," or "Cars with reported accidents have a significantly lower median price.")

Part 4: Dashboard Development in Tableau (25 Marks)

1. **(5 Marks)** Load the **fully cleaned and prepared** dataset (from Part 2) into Tableau.
2. **(15 Marks)** Create an interactive dashboard with at least the following visuals:
   - **KPI Cards:** Show the *Average Price*, *Average Mileage*, and *Total Cars Listed*.
   - **Bar Chart:** Show the average `price` by `brand` (for the top 10 brands).
   - **Pie Chart:** Show the proportion of cars that have a reported accident vs. those with a clean history.
   - **Scatter Plot:** An interactive plot showing `price` vs. `car_age`.
   - **Filters:** Add filters for `brand`, `fuel_type`, and `accident`.
3. **(5 Marks)** Organize your visuals into a professional, easy-to-read dashboard. Add a title and a brief summary explaining how a dealership manager could use it.

Submission Instructions

You are required to submit a single zip file containing:

1. Your Python notebook (`.ipynb`).
2. Your Tableau Workbook (`.twbx`).
3. A final PDF report with screenshots and explanations.