Skills For Hire Atlantic Cohort 6

Data Analytics - Assignment 1

Overview

This assignment consists of two parts: SQL Queries (Part 1) and Python Programming (Part 2) with a **maximum assignment credits of 25**. Answer both parts and provide clear explanations where necessary. You will be graded on the accuracy of your results. Please make sure to clearly indicate the question parts before answering them. The marks for each problem are indicated between brackets.

Note: This assignment will count for 60% of your total assignment grade.

Due date: May 4, 2025

Submission Instructions

1. **Compile Your Work**: Gather all your code/queries, outputs, and any explanations into a single document to ensure a well-structured submission.
2. **Screenshot & Explanation**: Include clear screenshots showing both the executed **code/query** and the **corresponding output**. Only output screenshots will not be accepted. You may also add comments or explanations to your code or results for clarity.
3. **File Format**: Save your document as a **PDF file (.pdf)** containing all your code/queries, outputs, and explanations.
4. **File Naming**: Name your file as `StudentId_FullName_Group.pdf` (e.g., `AM07677_AdeMartins_DA-Green.pdf`). Be sure to replace:
    - `StudentId` with your Student ID
    - `FullName` with your full name
    - `Group` with your TA group
5. **Submission Link:** Submit your file in the Assignment 1 box under **Curriculum** in the Community Learning Space (DISCO). You can follow Learning > *Your Group Name* > Curriculum, then scroll down to Assignment 1.

Reminder: Submitting a well-organized document will make it easier to review your work. Make sure everything is clear and complete before submitting!

Hints

- Start early. There are many parts to this assignment and it would be very difficult if left to the last minute.

- Don’t reinvent the wheel. Feel free to use the examples covered in class. If you get stuck, reach out to your TA for help!
- Do not spend hours without asking for help. Good luck!

Part 1: SQL Queries [10 points]

The first part contains five SQL queries that you need to answer. You are required to use the provided tables (`Customers`, `Orders`, `Shippings`) from the specified site (https://www.programiz.com/sql/online-compiler/).

Use the above tables to construct and execute the queries.

1) Find customers whose first name starts with 'J’, contains at least one character in between, and ends with 'n'. *The output query should only contain all details from the Customers table.* [1 point]

2) Find all orders placed by a customer with `customer_id = 2`. *The Output query should only contain all details from the Orders table.* [1 point]

3) Retrieve the first and last names of all customers who have placed an order with an amount greater than 500. *The output query should only contain the first name(s) and last name(s).* [2 points]

4) Show the total amount spent by each customer along with their first and last names. *The output query should only contain the first name(s) , last name(s) and total money spent.* [3 points]

5) Display all orders with their shipping status where the shipping status is 'Pending’. *The output should contain all order details and the shipping status.* [3 points]

Part 2: Python Programming [15 points]:

In this assignment, you will analyze a dataset containing traffic accident records to identify key trends, risk factors, and accident severity. The dataset includes information on crash dates, lighting conditions, crash types, primary causes, number of vehicles involved, injury severity, and accident timing. You will clean the data, explore accident trends, and visualize patterns to gain meaningful insights.

You are **required** to use **Python** and to **answer in a notebook**. You may use **Pandas**, **Matplotlib**, and **Seaborn** to process the dataset, create visualizations, and present your findings.

1. Loading the Dataset (1 mark)
a. Load the dataset `traffic_accidents_c6.csv` (attached to the Assignment on DISCO). **(0.5 mark)**
b. Display the first **10** rows. **(0.5 mark)**

2. Data Cleaning and Manipulation (3 marks)
a. Check for missing values and handle them appropriately. **(1 mark)**
b. Create a new column `crash_month` to store the accident month as a name (January, February, etc.). **(1 mark)**
c. Create a new column `crash_day` to store the accident day as a name (Monday, Tuesday, etc.). **(1 mark)**

3. Data Summary (3 marks)
a. Identify and list categorical columns. **(1 mark)**
b. Count the number of unique values for each categorical variable. **(1 mark)**
c. Generate descriptive statistics for numerical features. **(1 mark)**

4. Time-Based Trends (4 marks)
a. Visualize the number of accidents per month using a line plot. **(1 mark)**
b. Visualize the number of accidents per day of the week using a bar chart. **(1 mark)**
c. Identify the most accident-prone hour of the day and visualize it using an appropriate plot. **(1 mark)**
d. Compare the number of accidents that occurred in daylight and in darkness using two lines on the same plot. **(1 mark)**

5. Crash Type and Injury Analysis (3 marks)
a. Identify the top 5 most common crash types and visualize them using a bar chart. **(1 mark)**
b. Identify the top 5 causes of accidents and visualize them using a bar chart. **(1 mark)**
c. Display the number of fatal injuries at each hour of the day using a bar chart. **(1 mark)**

6. Heatmap Analysis (1 mark)
a. Create a heatmap to analyze accident occurrences across different hours of the day and days of the week. **(1 mark)**